VALIDATING USE OF A SYMPTOM ASSESSMENT SCALE IN PALLIATIVE CARE USING AN ARGUMENT-BASED APPROACH

by

Elayne E. Livote

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2011

This manuscript has been read and accepted for the
Graduate Faculty in Educational Psychology in satisfaction of the
dissertation requirements for the degree of Doctor of Philosophy.

Jay Verkuilen, Ph.D.

Date — Chair of Examining Committee

Mario Kelly, Ed.D.

Date — Executive Officer

Howard Everson, Ph.D.

David Rindskopf, Ph.D.

Ying Liu, Ph.D

Keith Markus, Ph.D.

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

VALIDATING USE OF A SYMPTOM ASSESSMENT SCALE
IN PALLIATIVE CARE
USING AN ARGUMENT-BASED APPROACH

by

Elayne E. Livote

Advisor: Professor Jay Verkuilen

Validation of patient-reported outcomes (PRO) scales has not kept up with contemporary views on validity and validation. For example, validity is not considered to be a binary state and it is the proposed use or interpretation of scale scores that is validated, not the scale itself. In this dissertation, I attempted to validate the use of a symptom assessment scale in a Veterans Affairs (VA)-based palliative care program to measure program outcomes using an argument-based approach to validity. In the first step of this approach, I developed the interpretive argument which specifies the claims and assumptions that are inherent in the proposed use. I then conducted three investigations to generate supporting evidence for the claims. The first was a basic psychometric analysis, the second was an assessment of measurement invariance, and the third was an examination of item directionality. In the validity evaluation, I assessed the plausibility of the claims incorporating the results of the investigations. I found that a bifactor model provided good fit to the data and concluded that while the psychometric properties

of the scale were fairly well maintained in this new use, the degree of missing data may be biasing outcomes and also prohibits use of the scale to measure outcomes. I also concluded that it may be more appropriate to treat some of the items of the scale as formative and this new formulation may help promote complete administration of the scale.

Acknowledgements

First and foremost, I would like to thank my dissertation advisor, Jay Verkuilen. It was both a pleasure and an honor to work with him. He gave me just the support I needed, when I needed it. I also benefited from the wealth of talented and helpful people in the psychometrics field I met through Jay.

I also want to thank my academic advisor and committee member, David Rindskopf. He is an excellent instructor and I learned so much from him. I also picked up valuable pointers on teaching that I hope to put into use some day.

I owe a big thanks to my other committee members for their input and insight, especially Howard Everson. The timing of his arrival at the GC and CASE was good fortune for me. I also want to thank Ying Liu and Keith Markus for their time and valuable comments.

I also want to acknowledge my classmates because they made it fun and were instrumental in getting me through the difficult times.

Thank you to my family and friends who thought I was nuts for doing this at my age, but nevertheless, were very supportive and encouraging, and genuinely happy for my success. I especially want to acknowledge my parents who instilled in me a love of learning.

The most important thank you goes to my life partner, llse de Veer. She not only encouraged me to do this and carried more than her share of the load these past six years, she also put up with the stress and late hours, especially while I was writing this dissertation. I am thrilled that I get to return the favor.

Table of Contents

List of Tables

List of Figures

Introduction

The NIH Patient Reported Outcomes Measurement Information System (PROMIS) Roadmap Initiative (www.nihpromis.org) is indicative of the growing importance and use of patient self-reported health status and quality of life in medical and heathcare research as well as in clinical practice. The goal of this program is to develop a large bank of items for measuring patient reported outcomes (PROs) and and to make these items available to researchers as individual items, short forms, and computer adaptive tests (Cella et al., 2007). A patient reported outcome refers to any patient self report of his or her quality of life (QOL), health-related quality of life (HR-QOL), health status, functional status, or symptom distress. There is growing recognition of the value of incorporating the patient perspective in clinical studies. The PRO Harmonization Group summarized the value of PROs as (a) providing a unique indicator of the impact of disease, (b) providing essential input for evaluating treatment efficacy, (c) assisting in interpreting clinical outcomes, and (d) a key element in making treatment decisions (Acquantdro et al., 2003). Furthermore, studies have shown that clinicians tend to understate patients' symptoms and the clinician reports are less sensitive to changes in patients status (Basch, 2010).

Substantial work has been completed on the PROMIS project including qualitative review and calibration of items in a number of domains, development of short forms, and the initiation of reliability and validity studies (Cella et al., 2010). Eleven item banks have been calibrated in areas such as physical functioning, pain, fatigue, sleep disturbance, depression, anxiety, and social health. But as item banks

become available to the public through the Assessment Center (www.assessmentcenter.net), users will be applying short forms in a variety of situations, creating new scales on the fly, developing scoring rules, and adapting items as needed, raising questions with respect to the validity of these new scales and new uses. Given the goals of PROMIS and the impact it is likely to have on the use of PROs, it seems an appropriate time to take a new look at validity and the validation process for health measurement scales.

This call to revisit validity and the validation process is motivated by current validity practice with respect to PROs and its implications for PROMIS, as well as the increased use of PROs to demonstrate value in health care. Measurement theorists have devoted much time and effort to validity and validation, and views have evolved substantially over time, but the validation of PROs has not kept up with these changes. The traditional view of validity is an accurate characterization of the approach still in use for PROs (Messick, 1989; Zumbo, 2005). According to this view, validity is a property of the measurement tool, and validity is a dichotomous state (that is, a scale is either validated or it is not). In addition, there are types of validity where the test user or developer assumes that only one type is needed, and validity is defined by a set of statistical procedures such as correlating scores with a criterion or gold standard (referred to as a validity coefficient). Finally, reliability is a necessary condition for validity.

Scales are routinely used for new purposes or in new circumstances without regard for validity because the user considers the scale as validated. This practice is contrary to the current perspective of validity which is that the interpretation,

inferences, or intended use of a test or scale is validated, not the test or scale itself (AERA, APA, & NCME, 1999). For instance, use of a scale to screen patients for post traumatic stress disorder (PTSD) requires different validation than the use of that same scale as an outcome in a study to compare PTSD treatments. This shift from the instrument to its use is the essential feature of the current approach with numerous implications. The first is that inferences are bounded by place, time, and use (Zumbo, 2005). Just as educational testing is used for a variety of purposes such as to assess student achievement, make admissions decisions, or evaluate schools, PROs are used as outcomes in studies, to screen for disease, or to evaluate a clinical program. This shift also makes the responsibility of scale users more explicit because the myriad purposes, populations, and modes of administration can not be anticipated by scale developers. A related point that is often overlooked is that the conditions under which a scale is developed are rarely found in real life applications.

Under the current perspective, validity is spoken of in terms of degree rather than as a binary state and it is viewed as the synthesis of evidence and theory in support of the proposed interpretation or inferences. The purpose is not to make validation an unattainable goal (Kane, 2001) but rather to instruct the user to weigh the existing validity evidence, much of which will presumably come from the validation activities of the developer with respect to the proposed use in order to identify where the evidence is lacking. It should be noted that the incremental evidence required may be minimal and need not be entirely empirical. It also may not be feasible to conduct a separate validity study in advance of applying the scale and some evidence may only be available after its use. Zumbo (2009) observed that

Otto Neurath's analogy of trying to build and rebuild a ship while at sea was also an appropriate metaphor for the concept of validity as well as for measurement validation. The important point is to emphasize the need to undertake the evaluation and to communicate the appropriate caveats or limitations on the conclusions that can be drawn.

It is also common practice in the validation of PROs to view validity in terms of type (see for example, Frost, Reeve, Liepa, Stauffer, & Hays, 2007) with a tendency to rely on one type. The trinitarian, or "three Cs", of validity are content, construct, and criterion. The current perspective is that validity is viewed as a unitary construct and types of validity have been replaced with types of validity evidence (AERA, APA, & NCME, 1999). This change might be appear to be purely semantics but it is consistent with the guiding principles of this new approach that the type of validity evidence should be driven by the proposed use and the nature of the construct and not merely selected on the basis of habit, convenience, or data availability. Also, moving away from types of validity serves the purpose of casting validity in a broader sense to include any aspect of the measurement process that could threaten the validity of the inferences made from test or scale scores (Cook & Cambell, 1979). In other words, using a valid scale does not ensure valid results or conclusions, because validity requires a use.

A validity-related issue that is rarely considered with respect to PROs is whether the construct is being measured by reflective or formative indicators, that is, are the items influenced by or do they influence the construct? This issue is paramount because it affects how items are selected and evaluated, reliability is

assessed, and the construct is validated. Because reflective measurement underlies *classical test theory*, factor analysis, item response theory (IRT), and IRT-based computer adaptive testing (CAT), scale items tend to be treated as reflective by default rather explicitly evaluated. However, there is growing awareness that many scales, including PRO scales, are inappropriately treated as reflective (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; Fayers & Hand, 1997; Streiner, 2003).

The issues raised above are not unique to PROMIS and apply any time an existing health measurement scale is developed or used, but the PROMIS project opens up a whole new realm of validity issues. To reiterate what was stated above, item banks will be tapped to create new scales on the fly, creating a hybrid situation where "validated" items will be combined in a variety of ways to create new scales obscuring the distinction between developer and user. There will also be a temptation or need to change the wording of items despite the fact that the PROMIS items underwent extensive qualitative review (Cella et al., 2010). For example, the recall or reference period may need to be adapted for the current situation. Likewise, users may want to change the number and wording of response options. These changes may or may not affect the psychometric properties, but they need to be investigated.

In the current health care environment, there has been and will continue to be an increased call for outcomes measurement and comparative effectiveness studies. The result is that more scales will be pressed into new uses for which they were not developed or validated. This use of scales raises additional validity issues such as applying the appropriate statistical methods in order to draw conclusions

about programs using measurements made at the individual level. A related validity challenge that is becoming more common is the use of scales for multiple purposes because the scales that are already in use for clinical purposes often become the source data for program evaluation.

To summarize, the view on validity and validation practice for PROs has lagged behind contemporary thinking in several ways. First is the notion that validity is a binary property of a scale, and a consequence of this notion is that scales are labeled as having been validated, and validity receives no further consideration. Another way that practice is out-of-date is the tendency to view validity in terms of types, specifically construct, content, and criterion. The problem is that each is treated as distinct, resulting in a validity process that is not necessarily comprehensive or coherent.

I agree with Zumbo's (2009) espousal of a meta theory of validity that transcends particular contexts (such as, education, cognitive, and health) but I also believe the biggest shortcoming of the current approach to the validation of PROs is that it fails to recognize things that are unique to PROs that threaten validity, and these things are not necessarily the traditional things that are thought of with respect to validity and validation. Perhaps the best example of this difference is with respect to missing data. In educational measurement, missing responses are usually treated as incorrect responses. Missing data in health measurement has been recognized as problem but not in the context of validity.

Goals of the Dissertation

The goals of this dissertation were to (1) assess the validity of the use of a symptom assessment scale that was adopted for use in a VA-based palliative care program to measure program outcomes, and in doing so, (2) provide a model for the validation of PRO scales, especially as it applies to the use of existing scales. The assessment included identifying evidence that is needed to support this use of the scale, conducting specific studies in an attempt to provide the evidence, and evaluating the extent to which use of the scale to measure outcomes is appropriate. I approached validation from the perspective of a user as opposed to a developer of a scale, but I also attempted to address issues that were not addressed in earlier validation studies, including those that are specific to its use in palliative care. As part of this process, I make very specific recommendations with respect to the validation of PROs.

To identify and evaluate the validity needs, I propose and demonstrate the use of an argument-based approach to validity (Kane, 1992, 2006; Kane, Crooks, & Cohen, 1999; Mislevy, Steinberg, & Almond, 2003; Shepard, 1993) which is based on Cronbach's (1988) validity argument. As Kane (1992) points out, justifying inferences made on the basis of an observed score implies the construction of an argument. Therefore this approach is consistent with and builds on the contemporary view on validity. I also believe that it is an especially useful approach for PROs because the intent is to provide a systematic way to identify and examine all the assumptions that are made when a scale is used for a particular purpose. Threats to validity are rooted in the plausibility of these assumptions.

Second, I suggest a greater role for measurement invariance studies to provide validity evidence when PRO scales are applied in new circumstances, uses, or conditions. Invariance studies are used to determine if the measurement properties of a scale are the same in two or more groups or circumstances. These studies are needed because the circumstances under which a scale is developed and validated are not generally found in real use. Also, as I demonstrate later, they could potentially address a number of validity issues at the same time, simplifying the validation process.

Developers and users of scales may not be aware that items of a scale can be of two types with respect to their relation to the construct. The distinction is whether the responses to the items of the scale are influenced by the construct or the items determine the construct (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). Reflective or effects indicators are so called because they reflect or manifest the effect of the construct and are by far the most common and usually default type. Formative or causal indicators form the construct (Fornell & Bookstein, 1982). In terms of the graphical representation, the difference is the direction of the arrows, that is, whether they point towards or away from the items to the construct. Appropriate treatment of scale items as reflective or formative has implications for development and validation of scales. The presence and therefore inappropriate treatment of formatively measured constructs seems more likely in health and quality of life measurement than in cognitive or psychological measurement. Therefore an examination of directionality should be a part of the validation process. Because there is a lack of examples in the PRO literature of cases where

directionality is examined in this context, I specifically address this issue and and demonstrate the use of an empirical method called tetrad analysis (Bollen & Ting, 1993) that can potentially be used for this purpose.

A recurring theme in this dissertation that has already been alluded to will be to emphasize and be more explicit about the roles and responsibilities of scale users with respect to validity. My goal is to convince scale users of the need to be more concerned with validity and to provide sufficient guidance on how to approach the seemingly daunting task of validating a particular use of a scale.

The organization of this dissertation is as follows. The unifying framework is the validation process formulated as a validity argument. The next section provides background information on palliative care and the program in which it is being used. This is followed by a methods section in which I explain the argument-based approach to validity and provide descriptions of the population, symptom assessment instrument, and general statistical methods. I then develop the validity argument with respect to the use of the scale to measure program outcomes. In the sections that follow, I carry out three investigations to provide support for the validity argument, each with its own methods, results, and discussion. The first is a basic psychometric analysis, the second is an assessment of measurement invariance, and the third is an examination of the directionality of the scale items. In the discussion section, I evaluate the validity argument and incorporate the results of the investigations followed by a comprehensive discussion.

Background

Palliative Care (PC)

Palliative Care is relatively new form of medical care that seeks to alleviate the burden and suffering associated with life-threatening diseases and their treatment including pain and other distressing symptoms (National Consensus Project for Quality Palliative Care, 2009). Though it is often associated with hospice and end-of-life care, palliative care can be delivered alongside curative care and is not necessarily the sole focus of care. In this sense, it helps patients gain the strength to carry on with daily life and improves their ability to tolerate medical treatments. Other goals of palliative care include assistance in making end-of-life decisions and attending to the spiritual and psychosocial needs of the patient.

Palliative care is operationalized through effective management of pain and other distressing symptoms. To that end, the National Consensus Project Clinical Practice Guidelines for Quality Palliative Care (2009) calls for the development of a timely care plan based on assessment of physical and psychological symptoms (Guideline 1.1). At the same time, the standards stipulate a commitment to quality assessment and performance improvement by implementing regular and systematic evaluation and measurement of processes of care and outcomes using validated instruments for data collection (Guideline 1.6).

Therefore, an important component of palliative care is the assessment of pain and other symptoms, but patients often lack the strength to complete lengthy assessments. The challenge is to obtain regular, timely, accurate, and comprehensive assessments of the patient's symptoms so that an appropriate treatment plan can be

developed, implemented, and evaluated with minimal burden on both the patient and the clinician. Several assessment scales have been developed for this purpose. For example, a list is provided under *Resources* by the National Palliative Care Resource Center (www.npcrc.org). The scales vary in their approach. Some are simple checklists of symptoms and others have attempted to measure not only the presence of symptoms but also their severity.

The VISN 3 Palliative Care Program

The Veterans Health Administration (VHA) provides health care to approximately 4.5 million enrolled veterans in 21 Veteran Integrated Service Networks (VISNs) across the United States (Sprague, 2004). In 2003, in conjunction with national initiatives to improve hospital care for patients with serious or advanced disease, the VHA directed all facilities to establish interdisciplinary palliative care consultation teams consisting of a physician, an advanced practice nurse, a social worker, a psychologist, and a chaplain (VHA, 2003). In response to this directive, VISN 3, which includes five acute care hospitals and three nursing homes in the New York City metropolitan area, developed and implemented a standardized, network-wide palliative care program with eight palliative care teams.

The VHA has been a pioneer in the use of electronic medical records (Jha, Perlin, Kizer, & Dudley, 2003), and as part of the standardized palliative care program, electronic templates were developed and adopted in October, 2004 to document consultations. The templates include the use of the Condensed Memorial Symptom Assessment Scale (CMSAS; Chang, Hwang, Kasimis, & Thaler, 2004) for symptom assessment along with a single-item categorical quality of life rating and

the clinician-rated Karnofsky performance status (Karnofsky & Burchenal, 1949).

Patients are assessed at the initial palliative care consultation in order to develop a care plan to address the most severe and problematic symptoms. The expectation is that the patient will be re-assessed within a short time (generally within 48 hours) to determine if the care plan is succeeding and to make any necessary adjustments. The scale is administered to the patient by a clinician (palliative care physician or advanced practice nurse). The scale can be self administered on paper, but it is not in this program. Data from the palliative care templates are extracted weekly from the electronic medical record to a database, where they are used as the source data for a report card containing program metrics (Penrod, Cortez, & Luhrs, 2007).

Methods

The Argument-Based Approach to Validity

In this section, I describe the argument-based approach to validity (Kane, 1992, 2006; Kane, Crooks, & Cohen, 1999; Mislevy, Steinberg, & Almond, 2003; Shepard, 1993), and explain why it is especially suitable to the validation of PROs. The term *argument-based approach* implies two things. First, it is a methodology for validation, not a type of validity. Kane (2006) notes that the treatment of validity as a unified construct, which is defined as the degree to which evidence and theory support the intended interpretation of test scores (AERA, APA, & NCME, 1999) is quite elegant but it is not a methodology. Most of the discussions regarding validity have been theoretical in nature or focused on specific types of validity (for example, construct validity), but have provided little guidance on the validation process itself. There has been a failure to address practical issues such as what types of evidence are required or how much evidence is sufficient. The argument-based approach is a framework for collecting and presenting validity evidence (Kane, 1992).

The other key term is *argument*. A score obtained from a test or a scale is of little use until some meaning is attached to it. It might correspond to some level of quality of life or it might be the used to determine who should be referred for treatment of depression. Validity is the extent to which the assigned meaning, interpretation or decisions made on the basis of the score are appropriate and justified. This definition of validity implies an argument, sometimes referred to as an interpretative argument, where the score is the premise and the interpretation is the conclusion (Kane, 1992). An argument-based approach to validity is akin to a

persuasive or informal argument (Toulmin, 1969) that is constructed to support the proposed interpretation. Unlike mathematical or logical proofs where assumptions are taken as given, the assumptions in informal arguments must be justified. Also, mathematical proofs are absolute. Informal arguments often involve qualifiers that indicate strength, or exceptions that identify circumstances in which the argument does not hold.

An argument-based approach to validity entails two sets of activities corresponding to two types of arguments. The first activity is formulating the interpretative argument, which begins with a statement of the proposed use or interpretation of a scale. The next step involves specifying the chain of inferences that are made in the course of administering a scale to the conclusions that are drawn or the decisions that are made based on the responses. The chain metaphor reinforces the idea that an argument is only as strong as its weakest link, therefore, validation efforts should be focused on the most questionable inferences and assumptions. Some of the inferences are subtle but are nevertheless important because they correspond to key aspects of measurement theory (Allen & Yen, 1989; Crocker & Algina, 1986). The final step is to identify the all the claims and assumptions upon which the inferences are made.

The validity argument or validity evaluation (Shepard, 1993) is an evaluation of the plausibility of the claims of the interpretative argument. Plausibility depends on the appropriateness and strength of the supporting evidence. This evidence can be empirical, analytical, or conceptual. Assumptions that remain doubtful may require further study. Kane, Crooks, and Cohen (1999) use a bridge analogy to

convey the idea that the inferences are a series of bridges that must be crossed. Wools (2008) takes this a step further by equating the supporting evidence as the tokens paid to cross the bridge.

Determining the required validity evidence based on an interpretive argument can guard against the tendency to rely the most easily obtained evidence (Kane, 1992; Shepard, 1993). At the same time, the particular circumstances will also dictate what evidence is available or feasible. A PRO scale may be used for the same exact purpose in two different circumstances but the interpretative and validity arguments will be very different because the set of assumptions and the amount, type, and availability of required evidence will be unique in each circumstance.

A large number of assumptions are made when a PRO scale is administered and an argument-based approach provides a systematic way to identify these assumptions so that they can be evaluated. Kane (1992) points out that the most problematic assumptions are those that are hidden because no supporting evidence can be provided if they have not been identified. An argument-based approach helps to uncover these hidden or unrecognized assumptions in the use of PROs.

The approach is not a radical departure from current practices in terms of the type of evidence. Rather, it is more about how the evidence is synthesized to form a coherent argument. Cronbach (1988) spoke of a *validity argument* rather than *validation research.* Likewise *The Standards* (AERA, APA, & NCME, 1999) advises researchers to develop a "sound validity argument [that] integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses"

(p.17). So the notion of a validity argument is not new, but in the absence of more specific guidance on how to put together such an argument, researchers simply created lists of of validity evidence. Chapelle, Enright, and Jamieson (2010) raised the question of whether this approach is really substantively different. They conclude it is and demonstrate how this approach overcomes the shortcomings of *The Standards* in the context of the TOEFL exam. For example, following *The Standards*, they generated a list of hypotheses emanating from the proposed use, but stopped at thirteen because they realized the list could be endless.

Study Sample

Two data sets were used for the analyses in this dissertation. The primary data set includes the symptom assessments for the patients who received a palliative care consultation between October 1, 2006 and September 30, 2009. This data set will be referred to as the palliative care or operations sample, and it contains 7,113 assessments from 1,886 subjects. A patient was counted as having had an assessment if at least one symptom was assessed. The second data set (referred to as the development or research sample) includes 924 subjects who participated in cancer related studies between 2000 and 2004. The subjects in this sample completed the Memorial Symptom Assessment Scale short form (MSAS-SF) and this sample was used to develop the CMSAS.

The characteristics of the samples are shown in Table 1. The subjects in both data sets are veterans who received health care at VA medical centers, therefore the sample consists almost entirely of older males. The mean age of the palliative care patients was 73.7 ($SD = 11.9$). The subjects in the research sample

Table 1

Characteristics of the Study Sample

Characteristic	Palliative Care	Research
Symptom Assessment Instrument	CMSAS	MSAS-SF
Purpose	Clinical	Research
Number of Subjects	1886	924
Number of Assessments	7113	924
Average Age (SD)	73.7 (11.9)	66.2 (11.2)
Male	1852 (98.2%)	910 (98.5%)
Primary Disease[a]		
Cancer	62.4%	100.0%
Chronic Obstructive Pulmonary Disease	4.3%	0.0%
Dementia	4.1%	0.0%
Congestive Heart Failure	3.9%	0.0%
Liver Disease	1.6%	0.0%
End-Stage Renal Disease	1.6%	0.0%
Other	22.1%	0.0%
Type of Facility		
Acute Care Hospital	88%	100%
VA Nursing Home	12%	0%

[a]Based on the disease recorded in the PC template as free text. Distribution based on non-missing, valid disease entries.

were generally younger with a mean age of 66.2 ($SD = 11.2$). The palliative care sample includes patients with a variety of underlying diseases but cancer is by far the most common (64.4%). The next most common diseases were chronic obstructive pulmonary disease (COPD) (4.3%), dementia (4.1%), congestive heart failure (CHF) (3.9%), end-stage renal disease (ESRD) (1.6%), and liver disease (1.6%). Another distinction between the samples is that 12% of the palliative care sample consisted of patients who resided in VA nursing homes.

Most of the subjects in the research data set had late stage (III or IV) disease and would have been considered eligible for palliative care. However, the cancer

studies preceded the implementation of the interdisciplinary palliative care program program so there should be negligible overlap of subjects. The most important difference between the two samples is that the subjects in the palliative care sample completed the assessments as part of their ongoing care. The subjects in the research sample were volunteers who consented to participate in research studies.

The Condensed Memorial Symptom Assessment Scale (CMSAS)

The CMSAS (Chang et al., 2004) is a scale that measures symptom distress. The instrument consists of 14 symptoms common in patients with life limiting diseases and includes 11 physical and 3 psychological symptoms. For the physical symptoms, the patient is asked if the symptom is present (yes or no) and if they respond affirmatively, then they are asked to rate the amount of distress or bother over the past seven days. There are five response categories: (1) not at all, (2) a little bit, (3) somewhat, (4) quite a bit, and (5) very much. If psychological symptoms are present, there are four response categories indicating how frequently the symptoms occurred in the past week. They are (1) rarely, (2) occasionally, (3) frequently, and (4) almost constantly. The additional response option for physical symptoms allows for symptoms that are present but are not causing any distress (Chang, Hwang, Feuerman, Kasimis, & Thaler, 2000). A copy of the CMSAS instrument is shown in the appendix.

The symptom distress score is calculated by assigning the values 0.8, 1.6, 2.4, 3.2, and 4.0 to the five physical symptom response options, respectively, and the values 1.0, 2.0, 3.0, and 4.0 to the four psychological symptom response options, respectively. A symptom that is not present is assigned a value of 0. The scale score

is calculated as the mean of the assigned values for the 14 symptoms. Physical distress and psychological distress subscale scores are similarly calculated by including only the relevant items. Therefore, symptom distress can range from 0.0 to 4.0. Chang et al. (2004) reported a mean score of 1.04 ($sd = 0.77$) with a range of 0 to 3.23 in the cancer population in which the scale was developed.

There are two earlier versions of this scale. The original scale is the Memorial Symptom Assessment Scale (Portenoy et al., 1994). It contained 32 symptoms measured with respect to three dimensions (severity, frequency, and bother/distress). A short form version of the MSAS was created in 2000 by Chang et al. by including only one of the three dimensions: bother/distress for the physical symptoms and frequency for the psychological symptoms. Thus the reference to symptoms as physical or psychological is based on the authors' classification. To create CMSAS, 18 symptoms were eliminated (17 physical and 1 psychological) on the basis of correlations with other clinical variables and significance in predicting survival (Chang et al., 2004).

General Methods and Software

SAS version 9.1 was used for data preparation and descriptive statistics. All models were run using Mplus version 6 as the primary software and results were confirmed using LISREL Version 8.8 with PRELIS version 2. Because the responses to the scale are ordered categorical, models were estimated using robust weighted least squares (WLS) (Muthén, 1984; Muthén, du Toit, & Spisic, 1997). The method involves a latent response variable formulation where the responses are treated as having come from an underlying continuous distribution that was divided into

categories at latent cutpoints. For example, if y is a dichotomous variable and y^* is the latent continuous variable, then

$$y = \begin{cases} 0 & \text{if } y^* \leq \tau_1, \\ 1 & \text{if } y^* > \tau_1 \end{cases} \tag{1}$$

where τ_i are the cutpoints for the latent continuous variable. They are called thresholds and are estimated parameters in the model. The raw data file must be used as input in order to estimate the thresholds. The raw data is also used to estimate the asymptotic covariance matrix which is used as the weight matrix in weighted least squares estimation.

The WLS method can be computationally intense with a large number of variables because the weight matrix is inverted. Robust WLS uses just the diagonal of the weight matrix to estimate parameters. The χ^2 statistic and standard errors are estimated using the full weight matrix but with with a method that does not involve inverting the weight matrix (Muthén, 1993; Muthén, du Toit, & Spisic, 1997). The χ^2 statistic is mean and variance adjusted, a method akin to the the Satorra-Bentler (1994, 2001) scaling factor. Degrees of freedom are not based on the number of parameters but instead are estimated to approximate a χ^2 distribution.

A large number of fit indices have been developed for structural equation modeling and factor analysis but only a few are appropriate for WLS estimation. In addition to the χ^2 test of model fit, Mplus reports the comparative fit index (CFI), the non-normed fit index (NNFI) which is also known as the Tucker Lewis Index (TLI), the root mean square error of approximation (RMSEA), and the weighted root mean square residual (WRMR) when robust WLS estimation is used. Hu and Bentler (1999) suggest reporting the standardized root mean square residual

(SRMR) plus one other index among CFI, TLI, and RMSEA. Mplus reports WRMR instead of SRMR but recently started recommending that it not be used citing its experimental nature (L. Muthén, 2010). The model χ^2 test is sensitive to sample size (Bentler & Bonnet, 1980), and will be included for comparison purposes but since it was always found to be significant in all the analyses it will not be discussed in reference to model fit.

The CFI (Bentler, 1990), a revised version of the normed fit index (NFI), compares the model being assessed to a more restrictive nested model which is typically the *null* or *independence* model in which all the covariances among the items are set to zero. It is more liberal than the absolute fit model χ^2 test and given the comparison model, it might suggest better fit than really exists. But it is also considered to be be well behaved and among the least affected by sample size (Fan, Thompson, & Wang, 1999). CFI values in excess of 0.95 are generally considered to indicate good fit. However, this benchmark is for continuous variables, and similar benchmarks for categorical or binary variables have not been established. In his dissertation, Yu (2002) conducted simulation studies to investigate cutoff criteria for categorical variables and concluded that values exceeding 0.96 should be considered good fit.

The TLI (Tucker & Lewis, 1973) also compares the current model to a null model. The TLI penalizes models for added complexity or additional parameters that do not improve fit. For categorical variables, Yu (2002) recommended values greater than 0.95 though this recommendation was less certain than the recommendation for CFI.

The Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) is often categorized as a measure of absolute rather than comparative fit even though it penalizes for lack of parsimony. It assesses the extent to which the model fits reasonably well in the population. When model fit is not perfect, the distribution of the fitting function F_{00} is *noncentral* χ^2 with a non-centrality parameter (NCP) that indicates the degree of shift of the χ^2 distribution, so that it also indicates the degree of model misspecification. NCP is estimated as $\chi^2 - df$. RMSEA adjusts for parsimony by dividing the normalized NCP by degrees of freedom. Values less than 0.05 were generally used as a cutoff for good fit though opinions are revising upward $(0.08 - 0.10)$. Yu (2002) recommends 0.05 as the cutoff for categorical variables. Mplus does not provide confidence limits or a p-value (the probability that the RMSEA < 0.05).

A diagonally WLS method is implemented in LISREL that produces parameter estimates that are almost identical to Mplus. The default goodness-of-fit indices are based on the Satorra-Bentler scaled χ^2 indicate better model fit than the corresponding Mplus statistics. Fit statistics based on Browne's (1984) asymptotically distribution free χ^2 can be obtained as an optional output file (Jöreskog, 2004) and are much closer to the Mplus fit statistics.

Analyses were carried out in the presence of missing data. Asparouhov and Muthén (2010) demonstrated that estimation with categorical outcomes using weighted least squares produces consistent and unbiased estimates using pairwise deletion under the MARX (missing at random with respect to covariates X) assumption. MARX is equivalent to MCAR (missing completely at random) when

there are no covariates in the model. Several models were rerun using full information maximum likelihood (FIML) estimation to confirm that results were consistent with the WLS results. WLS was preferred because of the time it took to estimate models using FIML.

Unless specifically noted, the IRT convention of freely estimating all loadings and fixing the factor variances to one was used. Other statistical methods used in the analyses are described in the sections in which they are used.

Developing the The Interpretative Argument

In this section, I develop the interpretative argument for the use of the CMSAS to measure outcomes in a VA-based palliative care program. It was previously mentioned that the scale is also used to guide clinical care which is its primary purpose, but I selected its use to measure program outcomes because it corresponds more closely to typical applications of PROs. However the fact that the scale is dual-purposed should be kept in mind as it has a direct bearing on the validity of its use as a program outcome.

One of the primary goals of palliative care is to improve quality of life through symptom relief. As a clinical metric, the Center to Advance Palliative Care (CAPC) recommends using the change in symptom distress over time (Weissman, Morrison, & Meier, 2010). This is operationalized as a change score calculated as the difference between the symptom distress score measured at the initial PC consultation and the score at re-assessment. The program outcome is the mean change score. Scores are calculated by facility and for the program as as a whole and is tracked over time.

In accordance with Kane's interpretative argument model, the following are the applicable inferences made from the administration of the scale to the interpretation or decision: (a) scoring, (b) generalization, (c) extrapolation, and (d) interpretation (Kane, 1992, 2001, 2006; Kane, Crooks, & Cohen, 1999). To the extent that these inferences are commonly found in interpretative arguments, the model helps identify all the relevant claims and organize the interpretative argument and will be employed here. Each inference will have a number of associated claims

and assumptions that are specific to the particular use and circumstances.

Scoring

This inference is from the observed behavior (responses to the scale) to the assignment of a score. Scoring is the process of assigning a numerical value to the set of responses to the test or scale. The basis for assigning a score can be objective or subjective, but both require the establishment of and adherence to a measurement procedure that defines how the observations are to be obtained. In the case of PROs, the measurement procedure is normally defined by the scale itself and its response options.

If scoring is subjective, such as the grading of an essay or the demonstration of a skill, the assumptions might be that there is an appropriate scoring rubric, inter-rater reliability among the scorers, and that the scorers have the appropriate credentials (Kane, 2006). PROs are generally objectively scored as either the sum or the mean of the responses or scores are derived from IRT models, especially when administered using computer adaptive testing. The assumption in the case of objective scoring is that the scoring rule is appropriate. Model-based scoring requires many additional assumptions with respect to model and person fit, and item calibration.

As previously described, the CMSAS score is a mean of the responses. Assigning numbers to ordered categorical response options and treating them as interval is common practice for PROs, but it is based on the assumption that the difference between consecutive options is constant across all the options. Also, a sum or mean score treats all symptoms as contributing equally to symptom distress.

Both of these assumptions are especially important in the context of measuring change in symptom distress.

Another very important assumption with respect the use of the scale to assess outcomes is that the score provide an accurate reflection of the patient's level of symptom distress insofar as a PRO should reflect only the respondent's input and the score for a patient should not depend on who administered the scale or how. Any number of scenarios could account for inaccurate scores. Because the scoring for the CMSAS is a mean, when the assessment is not administered completely, the level of distress for the missing symptoms is essentially assumed to equal to the mean level of distress for all the other symptoms. Given the prevalence of symptoms (see Table 5), it may be more likely that a symptom is not present so the mean score would be higher than the true level of distress. The opposite may occur if an informal approach is employed where any symptom not specifically mentioned by the patient is recorded as not present, that is, scores are understated. Clinicians may choose to focus only on the symptoms they perceive to be the most distressing or treatable, or they may simply be pressed for time. Missing symptoms or incorrect scores may also be a byproduct of the fact that despite the VA's use of electronic medical records, computers are not bedside and responses must be entered at a later time, so clinicians may only record the symptoms that are present.

Incomplete assessments may also reflect patient limitations. Patient burden is an important consideration for PROs. The patients in this population are elderly and have advanced disease. They are also likely to be limited by the very symptoms assessed by the scale, or by the treatments they receive for their disease and

symptoms. Another assumption concerns the motivation of the respondent. The scale is administered in the context of patient care where the relationship between the patient and clinician may be a source of response biases (Mazor, Clauser, Field, Yood, & Gurwitz, 2002). For example, patients may understate symptoms to please clinicians or overstate symptoms to get attention or drugs.

Generalization

The observed scale score is based on one instance of the measurement procedure taken under one particular set of circumstances. According to Kane's model, the generalization inference is what allows us to treat a single observed score as an expected score (1992, 2001, 2006). This claim can be made if the errors are sufficiently small and if the items included in the measurement are representative of the universe of possible items. The latter claim pertains to situations like educational testing where there is an almost limitless number of possible items.

The first claim is based on the assumption that tests or scales can be made sufficiently long to minimize random error. This also assumes that the construct is reflectively measured and thus items are considered to be exchangeable. However, limitations are placed on the lengths of PRO scales. One limitation is imposed by the subject matter, which is often easily exhausted. Another limitation is respondent burden, especially among the sickest populations. Computer adaptive tests attempt to get around this problem by only administering items that provide the most information in the vicinity of the respondent's estimated trait level. However, it is more often the case that scales cannot be as long as they ideally should be to reduce random error. In the present case, the CMSAS is the shortened

version of two previous scales, the MSAS (Portenoy et al., 1994) and the MSAS-SF (Chang et al., 2000). Given these limitations, the claim with respect to the CMSAS and PROs in general is that the scale is sufficiently free of measurement error. Zumbo and Rupp (2004) prefer the term precision over reliability because the meaning of reliability is ambiguous and they deem reliability as a data quality issue whereas validity is an inference quality issue.

The original scale, the MSAS, was developed for a cancer population and even though the many symptoms were eliminated from the CMSAS that could be thought of as more relevant to a cancer population (for example, hair loss), it was developed using a cancer population. Also, the fact that the conditions during development are rarely replicated in actual use has prompted some (Frost, Reeve, Liepa, Stauffer, & Hays, 2007; Zumbo & Rupp, 2004) to call for reporting reliability in all studies in which the scale is used and not just in validity studies. Therefore a related but additional claim with respect to the generalization inference is that the scale generalizes or is invariant across contexts. Use of the scale in a palliative care program differs from the development context in that it is being used in clinical operations and even though cancer is still the predominant disease, patients have variety of diseases such as congestive heart failure, dementia, and AIDs.

To summarize, the first claim is that a single measurement, which is the administration of the scale to a patient, is roughly equal to the expected score for that patient over hypothetical multiple administrations. In other words, the scale is reliable in the *classical test theory* sense. Secondly, because use of the scale involves aggregating scores over multiple patients, the other assumption that the scale

generalizes over the contexts in which it can vary in this program.

Extrapolation

The generalization inference makes a claim about the relationship between the observed score and the true score to use the *classical test theory* terminology, whereas the extrapolation inference is about the similarity between the scale or test to the target construct. In the present context, extrapolation is the inference made about the relationship between the construct of symptom distress as measured by this particular scale and the broader target construct which is the distress associated with the symptoms that are likely to be present in patients with life-limiting illness. The claims associated with this inference are that there is no construct underrepresentation or any construct irrelevant variance. Construct underrepresentation occurs if important symptoms are excluded from the scale or aspects of distress are not captured by the scale. Construct irrelevant variance refers to scores that are biased because they reflect something that is not related to the construct.

Decisions, Interpretation, or Actions

The decision inference refers to the actual conclusions, decisions, actions, or interpretations that are made on the basis of the symptom distress score. As indicated, the scores for individual patients are measured at the initial palliative care consultation and upon reassessment, they are used to calculate program metrics with respect to the program goal of symptom management.

The first claim is that the scale is sensitive enough to detect clinically meaningful changes in distress. Perhaps the biggest challenge with respect to PROs

Table 2

Claims in the Interpretative Argument for the CMSAS

Claims	
1.	The scoring rule is appropriate with respect to: equal weighting of symptoms the number of response options
2.	The observed score is free of bias or inaccuracies introduced by either the clinician or the patient
3.	The scale is relatively free of measurement error
4.	The scale generalizes across contexts
5.	No construct under-representation or irrelevance
6.	The scale is sensitive enough to detect clinically meaningful changes
7.	The scale longitudinally invariant
8.	Improvement can be attributed to palliative care

is determining the optimal number of response options. Additional options increase the responsiveness to change (Hays & Hadorn, 1992), but if the level of distinction is too fine, more burden is placed on the patient and they may find it difficult to respond. Another claim is that the scale is longitudinally invariant, meaning that the measurement properties remain consistent over repeated administrations.

The final claim is that improvements in symptom distress can be attributed to the palliative care treatment. It should be noted that this assumption is totally independent of the scale. Nevertheless, if this assumption does not hold then the conclusions drawn about the palliative care program will be invalid, reinforcing the notion that it is not the scale itself but the proposed use or interpretation of the scale that must be validated.

Table 2 provides a summary of the claims that make up the interpretative argument for the use of the CMSAS to measure program outcomes as identified above. The interpretative argument was developed without regard to what evidence

is required, or how that evidence may be obtained. The analyses found in the next three sections were designed to generate evidence to support or refute these claims.

Information from the model developed in the psychometric analysis such as the factor scores, factor loadings, and item thresholds will be used to evaluate the scoring rule. Symptom prevalence and missing data statistics generated as part of the psychometric analysis will supply evidence related to the claim that scoring is free of bias. The purpose of the psychometric analysis is to find a well fitting factor analytic model consistent with the expected factor structure of the scale. The degree of fit of this model to the palliative care sample will provide evidence of the claim that the scale is a precise measure of symptom distress. The invariance study will assess whether the measurement properties of the scale are preserved over all the contexts that can vary within the palliative care program such as underlying disease, palliative care team, patient setting, and time, providing evidence for the claim that the scale generalizes across contexts, and the claim of longitudinal invariance. Evaluation of the claim of no construct under-representation depends in part on whether the items of the CMSAS should be treated as reflective or formative measures of the symptom distress construct. This issue is addressed in the directionality study which will also provide additional insight when evaluating many of the other claims. Scale scores and factor scores from the model developed in the psychometric analysis will be used to assess the sensitivity claim.

It should be noted the analyses described above reflect the analyses that were intended but were not necessarily carried out. As with many studies, conditions may not be known until the study is underway. The analyses were also

not intended to be the sole source of evidence.

I return to the interpretative argument in the Discussion section where the plausibility of the argument is evaluated based on the results of the analyses, the strength of the studies, and other evidence. This is the validity evaluation which is the second part of the argument-based approach to validity.

Psychometric Analysis of the CMSAS

In this section, a basic psychometric analysis is carried out to assess the dimensionality of the scale in order to fit a factor analytic model to the palliative care and research samples. If appropriate models that fit the data can be found, then these models will also be used in subsequent analyses to assess: (a) measurement invariance and (b) whether the symptom distress construct should be reflectively or formatively measured.

Analysis of Missing Data and Symptom Prevalence

The psychometric analysis began with an examination of missing data in the palliative care sample and statistics related to symptom prevalence. Given that the scale employs a conditional format where the patient is first asked if a symptom is present (yes or no) and then they are asked to rate severity only if the symptom is present, missing data statistics were generated with respect to each part of the question. If a symptom is indicated as being absent, then the severity part of the question is automatically considered complete. Therefore, symptom prevalence affects completion statistics and was also examined. Prevalence in the palliative care sample was also compared to the research sample to provide an idea of the possible effects of missing data. Missing data was not analyzed in the research sample as it was almost 100% complete.

Missing data. Table 3 is shows the completion rates in terms of the number of symptoms assessed for presence or absence and severity where a totally complete assessment is one in which all 14 symptoms are assessed for both components. The results show that 24% of initial assessments were fully complete.

Table 3

Initial Symptom Assessment Completion Rates (Number of Symptoms Assessed) for the Palliative Care Sample

Number of Symptoms Assessed	Assessed For Presence/Absence (%)	Assessed for Presence/Absence and Severity (%)
14	41.8	24.0
13	35.4	25.9
12	6.0	8.2
11	5.1	6.6
10	1.2	5.9
9	1.0	5.1
8	1.1	5.3
7	0.9	4.5
6	0.9	3.6
5	1.3	2.9
4	1.3	2.2
3	1.9	2.6
2	1.0	1.1
1	1.4	1.5
0	0.0	0.7

If only presence or absence is considered then 42% are fully complete and the overall results confirm the suspicion that the data are much more complete with respect to the presence or absence component. Although not shown, completion rates for subsequent assessments are lower than for the initial assessments.

Completion rates were also examined by symptom and the results are shown in Table 4. With the exception of "Shortness of Breath" which was assessed for presence or absence 58% of the time, all the other symptoms were assessed approximately 90% of the time. As expected, the symptom that was assessed for presence or absence the most was "Pain" (96%) but it was not the most assessed for severity. Completion rates for symptom severity were calculated based on the symptoms that were present. Overall, severity was assessed 65% of the time when

Table 4

Completion Rates by Symptom for the Initial Assessment in the Palliative Care Sample

	Assessed for Presence or Absence	Assessed for Severity When Present	
Symptom	n (%)	Symptom Present (n)	Assessed n (%)
Lack of Energy	1,770 (93.8)	1,276	815 (63.9)
Lack of Appetite	1,757 (93.2)	893	570 (63.8)
Pain	1,819 (96.4)	975	651 (66.8)
Dry Mouth	1,693 (89.8)	569	403 (70.8)
Weight Loss	1,703 (90.3)	936	534 (57.1)
Feeling Drowsy	1,697 (90.0)	383	219 (57.2)
Constipation	1,725 (91.5)	405	248 (61.2)
Difficulty Sleeping	1,722 (91.3)	560	334 (59.6)
Difficulty Concentrating	1,690 (89.6)	428	267 (62.4)
Shortness of Breath	1,094 (58.0)	440	259 (58.9)
Nausea	1,722 (91.3)	234	145 (62.0)
Worrying	1,662 (88.1)	569	435 (76.4)
Feeling Sad	1,639 (86.9)	484	405 (83.7)
Feeling Nervous	1,567 (83.1)	336	270 (80.4)

the symptom was present. Completion rates for severity were more variable, and they were highest for the psychological symptoms, raising speculation that frequency may be easier to rate than bother or distress. Examination of missing data patterns did not reveal a tendency toward any particular pattern such as completion rates dropping off towards the end of the scale. Of course this assumes that symptoms are asked in the order they appear in the instrument.

Prevalence of symptoms. The prevalence of symptoms in the two samples are shown in Table 5. With the exception of "Lack of Appetite" and "Weight Loss", symptoms were more prevalent in the research sample. In both groups "Lack of Energy" was the most prevalent (72% in the palliative care sample

Table 5

Prevalence of Symptoms at Initial Assessment

Domain	Symptom	Palliative Care (%)	Research (%)
Physical	Lack of Energy	72.1	76.1
	Lack of Appetite	50.8	48.6
	Pain	53.6	72.8
	Dry Mouth	33.6	60.3
	Weight Loss	55.0	45.2
	Feeling Drowsy	22.6	54.4
	Constipation	23.5	38.4
	Difficulty Sleeping	32.5	52.5
	Difficulty Concentrating	25.3	33.9
	Shortness of Breath	40.2	53.9
	Nausea	13.6	28.6
Psychological	Worrying	34.2	44.5
	Feeling Sad	29.5	40.3
	Feeling Nervous	21.4	35.8

Note. Prevalence refers to symptom presence. Denominator equals the number of times the symptom was assessed in each sample.

and 76% in the research sample) and "Nausea" was the least prevalent (13% in the palliative care sample and 29% in the research sample). For symptoms such as "Feeling Drowsy" (54% vs. 23%), "Dry Mouth" (60% vs. 34%), and "Pain" (72% vs. 54%), the difference between the groups was substantial.

The prevalence rates show that the research sample experienced more symptoms than the palliative care sample and this finding was confirmed by examining the distribution of the number of prevalent symptoms. The median number of symptoms present in the research sample was 7 (IQR $= 4 - 10$) but only 4 (IQR $= 2 - 7$) in the palliative care sample.

Fitting Models to Both Samples Separately

Models were fit to both the palliative care research samples in anticipation of assessing measurement equivalence in the next section. For the present analysis, the palliative care sample was restricted to patients' initial assessment and only included patients whose primary disease was identified as cancer in the palliative care medical record template. The resulting sample size was 886. The restriction was made in order to eliminate potential sources of variation that will be examined separately either later in this section or at a later time. Also, as shown earlier, the primary source of missing data in the palliative care sample was the conditional component of the item with respect to severity that only applies if the symptom is present. As a result, the decision was made to perform the analysis using only the dichotomous component of the response that indicates whether the symptom is present or absent. The full research sample was used in this analysis ($n = 924$), but to be consistent only the binary responses to the symptoms in common with the CMSAS were used.

Dimensionality and the bifactor model. Tests and scales are generally designed to measure a single construct and therefore demonstration of unidimensionality is often offered as evidence of construct validity. To the extent possible, items are eliminated from scales until unidimensionality is achieved. The term *essential unidimensionality* (Stout, 1987) is used to describe situations where a scale is considered unidimensional enough in the sense that the scores are not substantially biased. Unidimensionality is also desired because it facilitates use of item response theory models and computer adaptive testing though this is becoming less of an obstacle (see for example, Gibbons, Immekus, & Bock, 2007; Reckase,

2009).

The bifactor model that has received a lot of attention lately, especially in the PRO literature (Holtzinger & Swineford, 1937). Initially, the model was employed as a method for assessing essential unidimensionality (Reise, Morizot, & Hays, 2007). The bifactor model is a hierarchical, multidimensional model with one primary or general factor and two or more item-group factors. All items have a non-zero loading on the general factor and non-zero loading on at most one of the item-group factors. The "bi" in bifactor describes the fact that items load on two factors and not that it is a two-factor model. In simplified terms, response data was deemed to have essential unidimensionality if the loadings for a one-factor model were not substantively different from the corresponding loadings on the general factor of the bifactor model.

The reason for the interest in the bifactor model was the growing recognition that health measurement scales are rarely unidimensional. For example, Cella and colleagues found two dimensions for self-reported health that they labeled as physical well being and mental well being (Cella, C.-H. Chang, Wright, Von Roenn, & Skeel, 2005). Often, the additional dimensions were attributable to the fact that scales were specifically designed to assess multiple domains of a construct (C.-H. Chang & Reeve, 2005). For example a scale that measures functioning might have subscales that focus on domains such as physical functioning, cognitive functioning, and social functioning. Based on similar reasoning, the bifactor has become a key model in the emerging testlet response theory literature (DeMars, 2006). I hypothesized a priori that a bifactor model was a good candidate model for the

CMSAS because it appeared to be closest to what the authors of the scale intended where the general factor corresponds to symptom distress and the item-group factors correspond to the physical and psychological domains.

Before fitting the bifactor model, exploratory factor analysis (EFA) was used as a first look at the factor structure. Table 6 shows the polychoric correlation matrix for the 14 items. The correlations were mostly in the moderate range $(0.2 - 0.7)$ but small correlations were found between "Shortness of Breath" with "Pain," "Constipation," and "Difficulty Concentrating". The scree plot of the eigenvalues corresponding to the binary responses to the CMSAS by the palliative care sample is shown in Figure 1. There are four eigenvalues greater than the often-used Kaiser-Guttman rule of 1.0, suggesting a model with more than one factor. However, the first eigenvalue is 5.6 and is more than three times the next eigenvalue of 1.6 indicating a strong common factor and possible essential unidimensionality. EFA models with one to four factors was examined. A maximum of four factors was selected because there are only 14 items. Fit did improve as the number of factors was increased but the pattern of factor loadings after geomin oblique rotations did not produce readily explainable factors in terms of common symptom clusters when the number of factors exceeded two.

Table 6

Polychoric Correlations

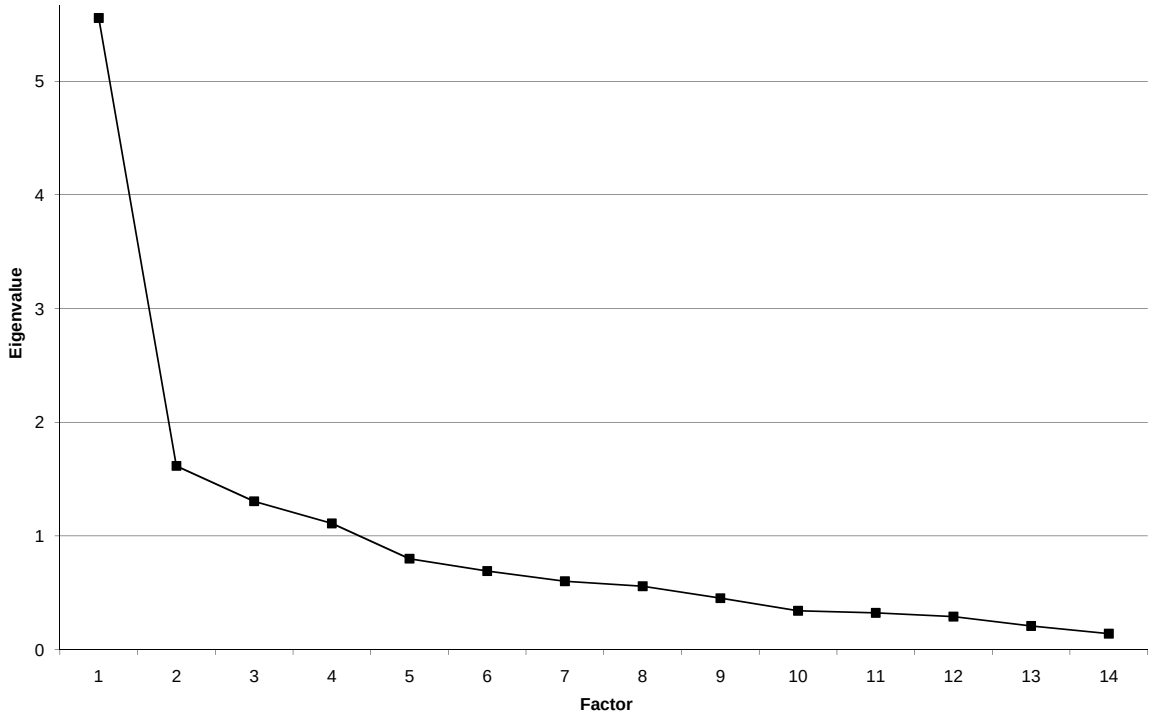
	Lack of Energy	Lack of Appetite	Pain	Dry Mouth	Weight Loss	Feeling Drowsy	Consti-pation	Difficulty Sleeping	Difficulty Concent	Shortness of Breath	Nausea	Feeling Worrying	Sad
Lack of Appetite	0.712												
Pain	0.348	0.243											
Dry Mouth	0.616	0.486	0.202										
Weight Loss	0.607	0.647	0.361	0.434									
Feeling Drowsy	0.569	0.468	0.316	0.494	0.267								
Constipation	0.219	0.272	0.393	0.232	0.316	0.232							
Difficulty Sleeping	0.420	0.308	0.384	0.310	0.234	0.290	0.287						
Difficulty Concentrating	0.511	0.321	0.109	0.420	0.174	0.475	0.142	0.349					
Shortness of Breath	0.404	0.444	0.034	0.320	0.286	0.233	0.010	0.164	0.030				
Nausea	0.363	0.425	0.300	0.146	0.327	0.218	0.309	0.326	0.173	0.213			
Worrying	0.453	0.227	0.387	0.238	0.345	0.268	0.196	0.486	0.303	0.242	0.319		
Feeling Sad	0.509	0.213	0.404	0.289	0.300	0.308	0.115	0.430	0.279	0.297	0.301	0.772	
Feeling Nervous	0.352	0.095	0.350	0.323	0.283	0.342	0.148	0.494	0.307	0.327	0.226	0.766	0.671

Based on the EFA results, one-factor and two-factor confirmatory factor models were run to serve as a comparison to the bifactor model. In the two-factor model, the physical items were constrained to load on one factor, the psychological items were constrained to load on the second factor, and the factors were allowed to have non-zero correlation. Global fit for the unidimensional model was fair with $CFI = 0.87$ and $RMSEA = 0.08$, but the two-factor model showed good fit ($CFI = 0.94$; $RMSEA = 0.06$). However, the strong correlation between the two factors (0.74) was consistent with the posited bifactor model.

Fitting the bifactor model. In a true bifactor solution, the general factor captures the variance shared among all the items, and the item-group factors capture the residual covariance among clusters of items. Therefore, all the factors should be orthogonal. Ideally, the loadings on the general factor should be salient and uniform. Figure 2 shows the bifactor representation of the CMSAS instrument as hypothesized. All 14 items load on the symptom distress primary factor and the first 11 items (the physical symptoms) simultaneously load on the physical item-group factor while the last 3 items (psychological symptoms) simultaneously load on the psychological item-group factor.

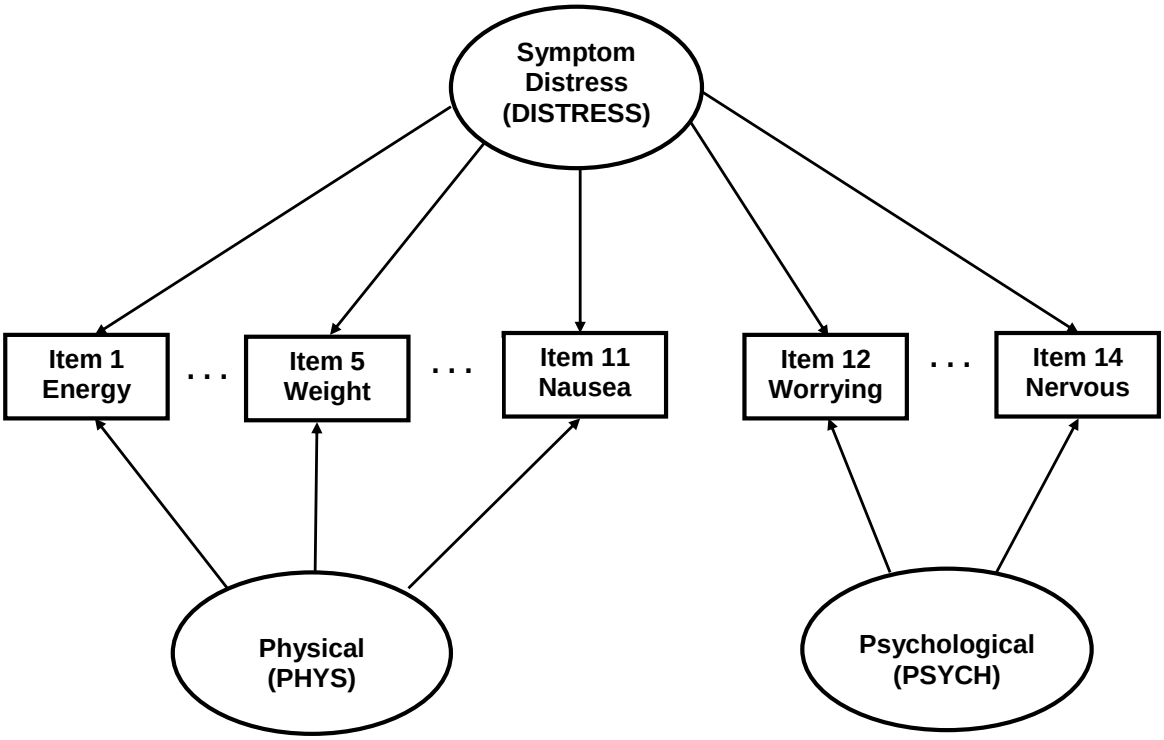
The bifactor model was run as described in the previous paragraph. The global fit of the model was good and exceeded the fit of the two-factor model (CFI= 0.97; TLI= 0.96; RMSEA= 0.04). The loadings are shown in Table 7. Three items had loadings below 0.4 on the general factor. "Shortness of Breath" had the lowest factor loading at 0.28, followed by "Lack of Appetite" at 0.32 and "Constipation" at 0.35. Even though a few loadings on the general factor were a

Figure 1. Scree plot of the eigenvalues for the binary responses to the CMSAS in the palliative care data set.



little below what what might be considered salient, none were near zero and all loadings were statistically significant ($p < .05$). For the physical item-group factor, the items "Pain" and "Difficulty Sleeping" have near zero loadings that were also not statistically significant but in a bifactor model it is not required that all items load on an item-group factor. However, according to my hypothesis, which was based on the design of the scale and its scoring rule, non-zero loadings were expected for all the physical symptom items. The three psychological items had loadings greater than 0.4 on psychological item-group factor. An examination of the modification indices suggested covariation between the uniquenesses for the pairs of items "Pain" and "Constipation," and "Difficulty Concentrating" and "Feeling

Figure 2. Bifactor representation of the CMSAS.



Drowsy" meaning that there is covariation between these pairs of items not accounted for in the general or item-group factors as hypothesized. Put another way, these items are not locally independent.

The bifactor model was also run for the research sample. The global fit of this model to the data was also good, and nearly identical to the fit in the palliative care sample (CFI= 0.97; TLI= 0.96; RMSEA= 0.04). The factor loadings are shown in Table 8. Like the palliative care sample, "Constipation" and "Lack of Appetite" had factor loadings below 0.4. A third symptom, "Weight Loss" had the smallest loading at 0.22. No loadings on the general factor were near zero and all were

Table 7

Factor Loadings for the Bifactor Model Fitted to the Palliative Care Data

	General	Item Group Factors	
Item	Symptom Distress	Physical	Psychological
Lack of Energy	0.632 (0.062)	0.620 (0.062)	
Lack of Appetite	0.321 (0.071)	0.848 (0.058)	
Pain	0.569 (0.051)	0.053 (0.070)	
Dry Mouth	0.431 (0.060)	0.488 (0.058)	
Weight Loss	0.420 (0.062)	0.549 (0.057)	
Feeling Drowsy	0.475 (0.060)	0.402 (0.068)	
Constipation	0.350 (0.057)	0.167 (0.065)	
Difficulty Sleeping	0.685 (0.051)	0.030 (0.075)	
Difficulty Concentrating	0.434 (0.058)	0.271 (0.066)	
Shortness of Breath	0.280 (0.076)	0.346 (0.078)	
Nausea	0.423 (0.064)	0.230 (0.071)	
Worrying	0.690 (0.053)		0.647 (0.080)
Feeling Sad	0.685 (0.055)		0.462 (0.079)
Feeling Nervous	0.649 (0.061)		0.490 (0.086)

Note. Fit indices: $\chi^2 = 168.6(63), CFI = 0.970, TLI = 0.957, RMSEA = 0.044$

significant. With respect to the physical item-group factor, the loading for "Difficulty Sleeping" was near zero and not significant which was also the case in the palliative care sample. "Difficulty Concentrating" was also near zero. So the pattern of factor loadings for the research sample was similar to the palliative care sample, but several items had very different loadings. Covariation between the uniquenesses for the pairs of items "'Pain" and "Constipation," "Nausea" and "Feeling Drowsy," and, "Lack of Appetite" and "Weight Loss" was identified through the modification indices.

Summary and Discussion of the Psychometric Analysis

In this section, a psychometric analysis was carried out that included an examination of missing data and a comparison of symptom prevalence in the

Table 8

Factor Loadings for the Bifactor Model Fitted to the Research Data

	General	Item Group Factors	
Item	Symptom Distress	Physical	Psychological
Lack of Energy	0.705 (0.049)	0.417 (0.073)	
Lack of Appetite	0.396 (0.065)	0.751 (0.066)	
Pain	0.571 (0.048)	0.241 (0.072)	
Dry Mouth	0.475 (0.048)	0.251 (0.066)	
Weight Loss	0.219 (0.062)	0.573 (0.063)	
Feeling Drowsy	0.614 (0.044)	0.263 (0.066)	
Constipation	0.381 (0.052)	0.344 (0.063)	
Difficulty Sleeping	0.581 (0.044)	−0.018 (0.072)	
Difficulty Concentrating	0.694 (0.042)	0.053 (0.075)	
Shortness of Breath	0.470 (0.046)	0.216 (0.065)	
Nausea	0.520 (0.051)	0.322 (0.068)	
Worrying	0.572 (0.043)		0.723 (0.078)
Feeling Sad	0.627 (0.043)		0.516 (0.068)
Feeling Nervous	0.593 (0.044)		0.380 (0.061)

Note. Fit indices: $\chi^2 = 168.8(63), CFI = 0.972, TLI = 0.960, RMSEA = 0.043$

palliative care and research samples. Assessments in the palliative care sample were rarely complete, especially the severity component of the question, while the research data was nearly complete. Symptom prevalence was generally lower in the palliative care sample but a comparison of symptom severity could not be made. These results suggest that very sick patients are capable of completing the scale even if it is only true of a subset of patients made up of mostly volunteers.

Factor models were fit to both samples using only the binary component of the items. After testing and rejecting one- and two-factor models, the posited bifactor model with item-group factors in accordance with the classification of symptoms as physical or psychological provided good fit to the data in both the palliative care and the research samples. Factor loadings for some items were not of

the magnitude consistent with a true bifactor model and correlations among the uniquenesses of some pairs of items was observed.

In the context of *essential unidimensionality*, the item-group factors are referred to as nuisance factors. Item-group factors are also referred to as method factors using the multi-trait multi-method terminology. In the case of the CMSAS, it is certainly possible that the item-group factors are the result of the fact that patients rate severity of physical symptoms in terms of bother or distress and they rate severity of psychological symptoms in terms of frequency. Though, it would be less likely given that only the binary responses were analyzed. A more likely explanation is that physical and psychological symptoms are qualitatively different, and one way they may be different is with respect to directionality. Regardless, the goal in this analysis was not to explain the true nature of the factors as much as it was to find a good fitting model that was justified based on the design of the scale. In that respect the bifactor is an appropriate model.

Residual correlation found among some of the physical items could also be indicative of the physical symptoms as formative indicators. In many respects, they behave like reflective indicators but since the factors could not account for all the covariation among these items suggests there is a different mechanism underlying responses to the scale. Fayers and Hand (1997) note that certain items may display especially strong correlations because they are clustered by common disease processes and treatments. Given that the scale was developed using cancer patients, and this analysis was conducted using only cancer patients, this explanation is certainly plausible.

Assessing Measurement Invariance

The second claim of the generalization inference of the interpretive argument was that the scale generalizes across contexts. The claim refers to the preservation of the measurement properties when use of the scale moves from the controlled research environment found during development to its application in a clinical program. It also refers to how well the measurement properties hold up in all the contexts that can vary within the palliative care program such as underlying disease, palliative care team, patient setting, and time. The latter is specific to the validity argument for use of the scale to measure program outcomes as it affects both precision in the form of variability, and the appropriateness of comparisons (for example, between palliative care teams). The former is a more general validity assessment that should be part of standard practice (Zumbo, 2005), but in the present case it also serves the purpose of allowing for an incremental approach to establishing generalizability across a number of contexts.

Background on Measurement Invariance

Measurement equivalence means that a scale measures the same construct in all groups and is a prerequisite for making comparisons between groups. A lack of invariance implies that the scale items fail to measure the construct in the same way in different situations (Meredith & Teresi, 2006). The central idea is that some of the measurement properties of the scale should be independent of characteristics of the person being assessed, other than the characteristic being measured by the scale (Millsap, 2007). This notion is captured in the formal definition used by Meredith (1993) and Millsap and Everson (1993) which states that measurement invariance

holds if

$$\Pr(X|W = w, V = v) = \Pr(X|W = w) \tag{2}$$

where X is a vector of scores on a measure, W is a vector of latent traits, and V is a scalar representing group membership. This definition is meant to be very general. X can be a continuous variable such as a test score or X can be a vector of responses to individual ordered categorical items. W can be unidimensional or multidimensional and group membership can be defined by observed or unobserved characteristics (see for example, Cohen & Bolt, 2005).

When responses to a scale are fit to a common factor model, measurement invariance is equated with factorial invariance though technically measurement invariance as implied by Equation 2 is stronger than factorial invariance (Millsap, 2007). Two methods of testing measurement equivalence using confirmatory factor analysis are multiple group CFA (Jöreskog, 1971), and multiple indicator multiple cause (MIMIC; Muthén, 1988) modeling. Both methods will be used as described below. Multiple group CFA was used to evaluate measurement equivalence between the research sample and the comparable palliative care sample (the initial assessments of cancer patients) and MIMIC modeling was used to assess measurement equivalence between the cancer and non-cancer palliative care subsamples.

Assessing Measurement Invariance Between the Research and Palliative Care Settings

Using multiple group CFA to assess measurement invariance.

Multiple group CFA involves simultaneously fitting CFA models in more than one

group. A prerequisite is that well-fitting parsimonious models have been found for each group separately (Brown, 2006, p. 271). Invariance is examined by placing equality constraints on the corresponding parameters in each group and assessing the deterioration in overall fit caused by the constraints. If the parameter is invariant, then there will be very little change in fit.

Most applications of factor analysis involve only the covariance structure:

$$\Sigma_{xx} = \Lambda_x \Phi \Lambda_x' + \Theta_\delta \tag{3}$$

where Σ_{xx} is the covariance matrix of the scale indicators, Λ_x is the matrix of factor loadings, Φ is the variance-covariance matrix of the latent factors, and Θ_δ is the matrix of residual variances-covariances. When the scale indicators are binary or ordered categorical variables, the the covariance matrix consists of polychoric correlations. However, measurement invariance also includes the parameters of the mean structure especially if group comparisons are going to be made. The factor model can be specified to include intercept parameters as follows:

$$X = \tau_x + \Lambda_x \Xi + \delta \tag{4}$$

where X is the vector of observed variables or item responses, τ_x is the vector of intercepts, Λ is the matrix of regression slopes, Ξ is the vector of latent factors, and δ is the vector of residuals. When specified this way the mean structure can be derived from (4) as

$$\mu_x = \tau_x + \Lambda_x \kappa \tag{5}$$

where μ_x is the vector of items means, and κ is the vector of factor means. For purposes of measurement invariance the applicable parameters are those of the

measurement model, specifically the factor loadings (Λ), intercepts (τ), and residual variances (Θ). When the scale responses are binary or ordered categorical variables, thresholds are also tested because they are parameters of the measurement model. Even though they are tied to intercepts, thresholds may not be invariant even if the intercepts are (Millsap & Yun-Tein, 2004). Equivalence of the factor means (κ) and factor variances and covariances (Φ) are considered tests of population heterogeneity.

Meredith (1993) described a hierarchy of measurement invariance which also suggests a step-wise procedure where increasingly restrictive equality constraints are imposed across groups in the multiple group CFA. This approach has more or less become the standard procedure (for example, Widaman & Reise, 1997; Brown, 2006, p. 269; Vandenberg & Lance, 2000) and was used in this analysis. The advantage of a step-wise approach is that is easier to identify the source if invariance does not hold.

At the lowest level, configural invariance describes the situation where the two groups have equal form, that is, the same number of factors and the same pattern of factor loadings. It would not make sense to test the equivalence of the parameters of the measurement model if configural invariance did not hold. In the subsequent steps equality constraints are placed on the corresponding parameters in each group. If model fit does not deteriorate significantly, then invariance for that parameter holds. Change in model fit is assessed by comparing nested models. The sequence then proceeds as follows:

1. Factor loadings or metric ($\Lambda^{(A)} = \Lambda^{(B)}$) $\Rightarrow$ Weak factorial invariance

2. Thresholds or scalar ($\tau^{(A)} = \tau^{(B)}$) $\Rightarrow$ Strong factorial invariance

3. Residual variances ($\Theta^{(A)} = \Theta^{(B)}$) $\Rightarrow$ Strict factorial invariance

Methods and sample for the multiple group CFA analysis. When robust WLS estimation is used, models can not be compared using the usual likelihood ratio χ^2 test because the difference in χ^2 values is not distributed as χ^2. Mplus has implemented a procedure they call DIFFTEST (Asparouhov & Muthén, 2006) that adjusts the difference between the mean- and variance-adjusted χ^2 statistics and calculates the appropriate degrees of freedom.

The samples used to assess measurement invariance between the palliative care and research contexts were the same as those used to fit the bifactor model in both groups. To reiterate, the palliative care sample consisted of the initial assessments for the 886 patients whose primary disease was identified in the palliative care template as cancer. The full research sample of 924 was used. In both samples, only the binary component of the response was used.

Results of the multiple group CFA analysis. The prior section demonstrated that the postulated bifactor model fit both the palliative care cancer and research samples well. These results are repeated at the top of Table 9 under the section labeled "Single Group". The results of testing invariance are also shown in this table. Configural invariance is not tested directly except by examining the fit indices when multiple-group CFA is run with the same factor structure for both groups but no equality constraints are imposed. The results in Table 9 show that fit is comparable to the fit found in the separate models ($CFI = 0.971$, $TLI = 0.956$, $RMSEA = 0.043$). The χ^2 statistics are included in this table to provide a sense of the relative fit but as previously noted, the overall fit statistic was always significant

Table 9

Tests of Measurement Invariance Between the PC and Research Samples Using Multiple Group CFA

Model	χ^2	*df*	CFI	TLI	RMSEA	χ^2	*df*	
Single Group								
Palliative Care	168.6	63	0.970	0.957	0.044			
Research	168.8	63	0.972	0.960	0.043			
Measurement Invariance								
Configural Invariance	337.3	126	0.971	0.956	0.043			
Metric	401.1	151	0.966	0.959	0.043	79.3	25	*
Partial Metric[a]	329.1	146	0.975	0.969	0.037	28.1	20	
Scalar	416.1	152	0.964	0.957	0.044	83.3	6	*
Partial Scalar[b]	328.9	150	0.976	0.970	0.036	2.5	4	
Equal residual variances	342.7	143	0.973	0.965	0.039	18.7	7	*
Partial residual variances[c]	335.0	142	0.974	0.966	0.039	12.1	6	

Note. The χ^2 statistic is the Mplus DIFFTEST χ^2.; CFI=comparative fit index; TLI=Tucker Lewis index; RMSEA=root mean square error of approximation.

[a]Non-invariant loadings: Weight Loss, Concentration, Pain, Constipation, & Nausea.

[b]Non-invariant thresholds: Difficulty Sleeping & Lack of Energy.

[c]Non-invariant residual variances: Shortness of Breath.

*$p < .01$.

which is attributable to the large sample size. Models are formally compared using the the DIFFTEST χ^2 statistic which is an adjusted difference in the overall χ^2 value.

Metric invariance was tested by constraining the corresponding factor loadings to be equal in both groups. The overall fit did not change by much ($CFI = 0.966$, $TLI = 0.959$, $RMSEA = 0.043$), and in fact TLI increases because fewer parameters are being estimated. But the DIFFTEST χ^2 statistic is significant (79.3 with $25df$). Modification indices were examined, and equality constraints were relaxed one loading at time until the DIFFTEST statistic was no longer significant.

Non-invariance was found for the loadings for "Weight Loss" and "Concentration" on the general factor and "Pain," "Constipation," and "Nausea" on the physical item-group factor.

Byrne, Shavelson, and Muthén (1989) introduced the concept of partial invariance in which they demonstrated that the testing procedure can continue as long as there is a subset of items where invariance holds. In some circumstances it still may be appropriate to compare factor means. Even though it is not certain if the current circumstances qualify, invariance testing for the remaining items continued using the process described above, but with the non-invariant parameters freely estimated. They also note the importance of verification using cross-validation to confirm the parameters identified as noninvariant. Unfortunately, there was not sufficient sample size to cross validate at this time.

The test for scalar invariance was significant (DIFFTEST $\chi^2 = 83.3$ with $6df$), but partial scalar invariance was achieved after freeing the thresholds for "Difficulty Sleeping" and "Lack of Energy." In the last step, equality of residual variances was tested for the remaining invariant items. "Shortness of Breath" was found to have non-invariant residual variance after a significant DIFFTEST χ^2 statistic (18.7 with $7df$).

Assessing Measurement Invariance Between the Cancer and Non-Cancer Palliative Care Subsamples

Using MIMIC modeling to assess measurement invariance.

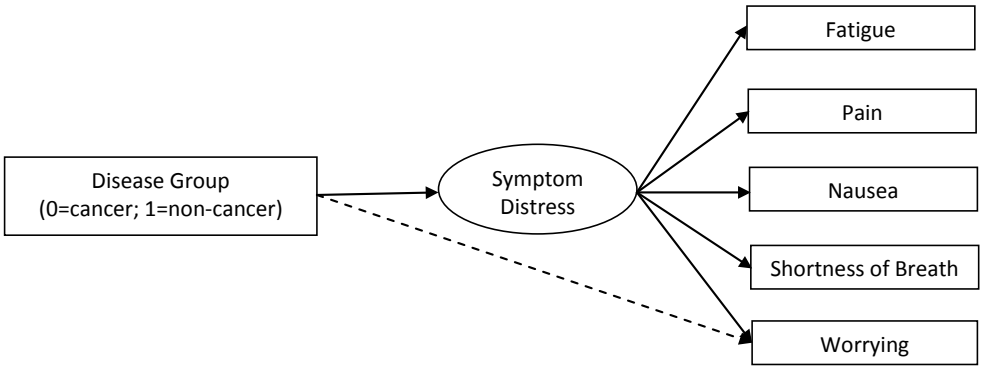
Implementation of Multiple group CFA requires large overall sample sizes with approximately equal sample size in each group to prevent one group from

overwhelming the other when assessing equality of model parameters. Therefore multiple group CFA could not be used to test invariance between the cancer and non-cancer palliative care populations because of the size of the non-cancer sample and because of the size imbalance. MIMIC modeling in which a covariate indicating group membership is added to a single-group CFA model was used instead. Fewer parameters are estimated in a single-group model, and, unlike multiple group CFA, MIMIC modeling does not include the estimation of intercepts and factor means. Fewer estimated parameters results in a smaller sample size requirement.

With the growing use of item response models and the ability of software such as Mplus and PRELIS/LISREL to handle categorical indicators, the lines between CFA and IRT have been blurred even though the similarities between the two have been noted for some time (for example, Bartholomew & Knott, 1987; Takane & de Leeuw, 1987). If in Equation 4, the vector X contains item scores such as in the present analysis, then measurement invariance refers to differential item functioning or DIF (Thissen, Steinberg, & Wainer, 1988).

In the MIMIC approach to assessing invariance, the model factors and items are regressed on the variable indicating group membership. A simplified version of the MIMIC model used in this analysis is depicted in Figure 3. It shows a solid regression line from the covariate for disease group to the symptom distress general factor and a dashed line directly to the item "Worrying." If the coefficient for the direct path from the covariate to the item is not significant, then any effect of the covariate is fully mediated through the latent trait. In other words, response differences between the groups for that item is fully explained by differences in the

Figure 3. Example of a MIMIC model to assess measurement invariance.



level of the latent trait. On the other hand, significance of the coefficient for the direct path indicates measurement non-invariance and the the coefficient can be interpreted as a measure of the magnitude of the DIF (Fleishman, 2004). If the regression path from the covariate to the latent trait is significant, then it indicates population heterogeneity, in other words, the the groups differ on their mean level of the latent trait.

It should be noted that in the language of DIF, MIMIC models can only detect uniform DIF. Uniform DIF means that the response bias is constant across the trait spectrum (Teresi, 2006). In terms of factorial invariance, MIMIC models test the equivalence of the intercepts (scalar invariance), but assume that the factor loadings are equal across groups and this assumption is not formally tested. Multiple group CFA can detect non-uniform DIF and provides a more robust assessment of measurement invariance.

The method is implemented in three main steps. In the first step, a

single-group CFA model is fit to a data set that contains both samples but without the group membership covariate. In the second step, the indirect effect of the covariate on item responses is added to the model as regression coefficients from the covariate to the factors. A technique that is used to determine if any items are non-invariant is to add the direct effects (the regression of the items on group membership) to the model but constrain the coefficients to be zero. The presence of the coefficients for the direct paths in the model will create modification indices for constrained parameters if fit can be improved by removing the constraint (Muthén, & Muthén, 2009). In the third step, the zero constraints are removed as suggested by the modification indices allowing the coefficients to be estimated and tested for statistical significance. Models can not be compared using χ^2 difference testing because the models are not nested.

Sample used for the MIMIC analysis. The non-cancer sample consists of palliative care patients whose primary disease was readily determined from the free text entry in the palliative care template but was not cancer. The sample size in this group was 247 and the most common diseases in this group were COPD (61), dementia (58), and CHF (56). Consistent with the prior analyses, only the binary component of the patient's initial assessment was used. The palliative care cancer sample was the one used in the multiple group CFA analysis to assess measurement equivalence between the palliative care and research contexts ($n = 886$).

Results of assessing measurement invariance across disease groups using MIMIC models. The results of the MIMIC analysis are shown in Table 10. The first row shows the global fit indices for the single-group model with no disease

Table 10

Tests of Measurement Invariance Between the Cancer and Non-Cancer Palliative Care Samples Using MIMIC CFA Models

Model	χ^2	*df*	CFI	TLI	RMSEA
Single Group, no disease group covariate	213.1	63	0.968	0.953	0.046
Indirect effects, factors regressed on disease group covariate	298.8	74	0.952	0.932	0.052
Add Direct effects - items[a] regressed on disease group covariate	214.7	71	0.969	0.954	0.042

Note. Disease group covariate: 0=cancer ($n = 886$), 1=non-cancer ($n = 247$).

[a]Pain, Weight Loss, and Shortness of Breath.

group covariate ($CFI = 0.968$, $TLI = 0.953$, $RMSEA = 0.046$). Not surprisingly, the fit indices are very similar to those obtained when the bifactor model was fit to the cancer palliative sample (see Table 7) since this population dominates the sample currently under consideration. The disease group covariate was included in the second and third models. The second model (second row) served the purpose of helping identify any non-invariant items by examining global fit and the modification indices for the group by item coefficients. The zero constraint was removed one item at a time until there were no longer any modification indices for the group by item coefficients. The items identified as non-invariant were "Pain," "Weight Loss," and "Shortness of Breath." The final model (third row) includes the significant group by item direct effects for these items and the global fit of this model is good ($CFI = 0.969$, $TLI = 0.954$, $RMSEA = 0.042$).

Table 11 shows the disease group by factor and disease group by item coefficients along with the standard errors and p values. Non-zero group by factor coefficients indicate population heterogeneity; the results show that the cancer and

Table 11

Disease Group Effects in the MIMIC Model

Regressed on Disease Group	Coefficient	SE	p
General Distress	−0.053	0.127	0.679
Physical Distress	−0.428	0.120	$< .001$
Psychological Distress	−0.177	0.192	0.356
Pain	−0.574	0.100	$< .001$
Weight Loss	−0.535	0.096	$< .001$
Shortness of Breath	0.401	0.128	0.002

non-cancer groups do not differ with respect to general symptom distress or psychological distress but they differ on physical distress. The coding for the covariate was 1 for the non-cancer group and 0 for cancer, so the negative coefficient indicates that the non-cancer group had less physical distress.

The interpretation of a significant group by item coefficient is that for a given level of the latent trait, the groups differ on how they respond to the item. This is the very definition of differential item functioning (DIF). The negative coefficients for "Pain" and "Weight Loss" mean that for a given level of distress (general, physical, and psychological), cancer patients will be more likely to experience pain and weight loss than the non-cancer patients and the positive coefficient for "Shortness of Breath" means that non-cancer patients are more likely to experience shortness of breath at the same level of distress.

Summary and discussion of the measurement invariance study. Multiple-group CFA was used to assess measurement invariance between the cancer patients in the palliative care sample and the research sample in which the scale was developed. The analysis in the previous section showed that the same bifactor

model provided good fit in both samples and the loadings on the general factor were comparable. But a closer examination of measurement equivalence revealed that more than half the items were not invariant with respect to one or more of the measurement parameters. Five items did not have invariant loadings indicating different relationships between the item and the latent variable. Loadings correspond to the discrimination parameter in IRT. Two items did not have invariant thresholds, which correspond to the difficulty parameter in IRT. One more item had non-invariant residual variance. It should be noted that the large sample size in both groups meant that the test was able to detect small degrees of non-invariance and that based on the changes in global fit, the non-invariance did not appear to be substantial.

Multiple-group CFA allows for a more thorough assessment of invariance than MIMIC modeling but MIMIC modeling can be used when the sample sizes are small and imbalanced. For these reasons MIMIC modeling was used to assess invariance between cancer and non-cancer subgroups of the palliative care sample. Three items were found to have DIF: "Pain," "Weight Loss," and "Shortness of Breath."

This analysis assessed invariance between the development context and a real life application of the scale. Invariance was also assessed within the palliative care population, specifically invariance between the cancer and non-cancer subgroups. Invariance assessment is by no means complete. Longitudinal invariance refers to the measurement properties of a scale under repeated administrations. It is a well recognized source of invariance (for example, Pentz, & Chou, 1994) that is

particularly relevant to the use of a scale to measure outcomes. However, it could not be assessed in this study because of a lack of data. Subsequent assessments, when attempted at all, were far less complete than the initial assessments.

Cohen and Bolt (2005) point out that while a myriad of methods exist for detecting non-invariance or DIF, they do not provide insight as to why members of one group respond differently to an item than than members of another group. The reason for this limitation is that non-invariance is detected on the basis of manifest characteristics (for example, gender). They suggest that the mechanism responsible for non-invariance is probably not perfectly synonymous with the group designation used to detect the DIF. They propose a strategy of identifying non-invariant latent classes using mixture modeling. This analysis will be carried out at a later time when there is additional data with sufficient diversity.

While it is not known why some items were non-invariant, it is interesting to note that the psychological items were invariant across the board in both invariance investigations. The MIMIC analysis also showed a non-significant indirect effect of disease group on the psychological item-group factor which indicates that there is no population (cancer and non-cancer) heterogeneity with respect to this factor. These results could be another indication that the physical and psychological symptoms are different.

Assessing Directionality of the CMSAS Items

In this investigation, I consider the directionality of the items that make up the CMSAS. Directionality refers to whether the responses to the items of the scale influence or are influenced by the latent construct. Observed variables that are specified as being caused by the latent variables are termed "effects" or "reflective" indicators (Edwards & Bagozzi, 2000) because they reflect or manifest the effect of the latent construct. Items that determine the construct are referred to as "causal" or "formative" indicators because the construct is formed by its measures (Fornell & Bookstein, 1982). The type of measurement has implications for scale development and validation.

Background on Formative and Reflective Measurement

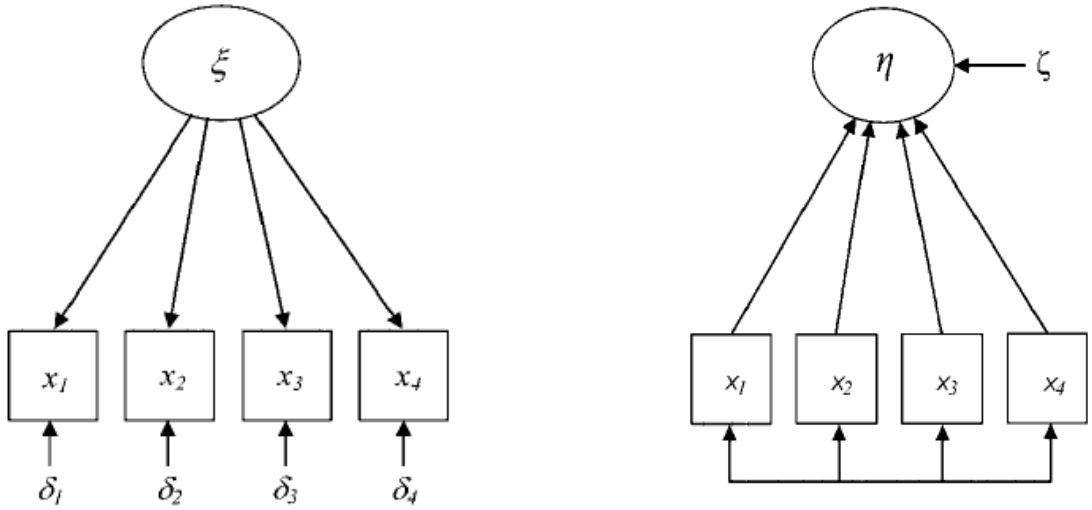
Reflective measurement underlies classical test theory (Lord & Novick, 1968), factor analysis (Spearman, 1904), and standard latent variable models (Bartholomew & Knott, 1999), including item response theory (Embretson & Reise, 2000) and is often appropriate for cognitive and psychological measurement. But PRO scales in general and the CMSAS in particular have been treated as reflectively measured *de facto* by virtue of how these scales were developed and validated rather than explicitly evaluated. Edwards and Bagozzi (2000) refer to it as a failure to develop an auxiliary theory that addresses the nature and direction of the relationship between a construct and its indicators. The issue is being raised in this context specifically because of the recognition that some variables in health and quality of life measurement determine the latent construct and therefore should not be treated as reflective (Boehmer & Luszczynska, 2006; Bollen & Lennox, 1991;

Fayers & Hand, 1997; Streiner, 2003; Zumbo, 2007).

Scales based on formative measurement are much more common in fields such as economics and business. The construct is often referred to as a latent composite or just a composite variable (Fayers & Hand, 1997), reflecting the notion that the measures included in the scale become the operational definition of the construct (Bagozzi, 1982). The most often cited example of formative measurement is socio-economic status (SES; Hauser & Goldberger, 1971). SES is measured by a combination of income, education, and occupation. A common example of a health-related formative measurement scale is life stress (Bollen & Lennox, 1971), where it is clear to see that each stressor is an independent contributor to the latent variable.

Figure 4 depicts two constructs. The construct on the left is measured by reflective indicators and the construct on the right is measured by formative indicators. The difference between the two models is more than just the direction of the arrows. In the reflective model the the latent variable is exogenous and errors are accounted for at the item level. There is no directly caused correlation between the items because covariation is captured by the latent variable. Residual variance represents the item variation that is not shared with or explained by the latent construct and serves as an indication of how well the item measures the construct. The arrows point from the latent variable to the indicators, implying that changes in the construct should produce changes in all the indicators. A corollary of reflective measurement is that items are considered to be exchangeable (Zumbo, 2007), which allows for things such as the development of item banks and IRT-based

Figure 4. Examples of reflective and formative indicator models.



computer adaptive testing.

Bollen and Lennox (1991) observe that classical test theory and factor analysis with its reflective-type measurement has greatly influenced beliefs about the qualities that define valid and reliable items. One such belief belief is that if items are positively associated with the construct, then they should be correlated with each other. Scale development procedures often include examination of item-scale correlations, and items displaying poor or negative correlations are dropped from the scale. Exploratory and confirmatory factor analytic methods are employed in the scale development process. Reliability is frequently assessed as internal consistency reliability (Fayers & Hand, 1997; Streiner, 2003). These procedures are followed without regard to whether items are are reflective or formative.

But an examination of the formatively measured construct on the right side of Figure 4 demonstrates why the usual approach to scale construction is not

appropriate when the indicators cause or induce the latent construct. In formative measurement, the arrows point towards the latent variable from the indicators indicating that each item contributes independently to the latent variable. A change in just one of the indicators will produce a change in the construct without requiring or inducing an accompanying change in the other indicators. Therefore, there is no reason to expect that the items should covary and in fact, highly correlated items can cause instability due to multicollinearity. Dropping items that do not correlate with the other items could substantially change the construct as that item's unique contribution is then lost. Each item can be considered as representing a distinct dimension of the construct and therefore the scale is multidimensional (Bollen & Lennox, 1991). This contrasts with reflective measurement where internal consistency is equated with unidimensionality. In other words, each latent variable is considered a unidimensional construct.

Formative scales are generally longer to ensure sufficient coverage of the construct domain to prevent construct underrepresentation. Because reflective items are considered exchangeable, the purpose of longer scales is to decrease sampling error. Formative measurement is designed to minimize residuals in the structural relationship (Fornell & Bookstein, 1982), so precision of formative scales is based on the size of the error term on the latent variable. But a scale made up of all formative indicators is not identified. A necessary but insufficient condition for identifying the error term is that the latent variable influence at least two other reflectively measured latent or observed variables (MacCallum & Browne, 1993). An example would be a hybrid latent variable that is measured by a mixture of

formative and reflective indicators. A mixed indicator model is another example of a MIMIC model.

Sum or mean scoring without weighting of items is often used for scales. This approach is appropriate for reflective scales because the items are parallel measures of latent variable. In formative measurement, each item is an independent dimension of the construct and should be weighted appropriately to reflect the magnitude of its contribution to the construct (Bucic & Gudergan, 2004).

Given the fact that many models involving formative items are underidentified, Bollen (1989, pp. 65-7) recommended the use of *mental experiments* to determine the correct specification of an indicator. The process is similar to the discussion above where a change in the construct is imagined and causal reasoning is applied to determine if a change in the item response is expected to follow. Bollen and Ting (2000) note that establishing a causal priority between a latent variable and its indicators can be difficult. Determining the appropriate treatment of symptoms can be especially challenging because they can be formative in some contexts and reflective others. Symptoms can also an have a reciprocal relationship with the latent variable meaning they can be both formative and reflective at the same time. For example, difficulty sleeping can lead to distress but distress can also be manifested in sleep disturbances.

Methods and Sample for the Confirmatory Tetrad Analysis

In this study, I used confirmatory tetrad analysis (CTA) which is an empirical method that was proposed by Bollen and Ting (1993; 2000) to assist in making the determination between reflective and formative indicators. A tetrad is

the difference between the product of a pair of covariances among four variables and the product among a different pair of the same variables. For example, using the notation attributed to Kelley by Bollen and Ting (1993), the tetrad τ_{1234} is $\sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24}$. Spearman (1904, 1927) demonstrated that in factor models some tetrads are expected to be equal to zero, which led to the term "vanishing tetrad." Prior to the development of other methods such as maximum likelihood and its implementation in LISREL, it was used to evaluate model fit. Glymour, Scheines, Spirtes, and Kelly (1987) proposed the use of tetrads for exploratory purposes by searching for models that were consistent with the observed tetrads.

The following development comes from Bollen and Ting (1993). Covariance algebra can be used to show that a latent variable that is measured by four reflective indicators has three vanishing tetrads:

$$\begin{aligned} \tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24} = 0 \\ \tau_{1342} &= \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32} = 0 \\ \tau_{1423} &= \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43} = 0 \end{aligned} \tag{6}$$

Bollen (1990) developed an asymptotically distribution-free simultaneous significance test that tests whether, given sampling error, the observed vanishing tetrads are consistent with the model implied vanishing tetrads and thus provides a goodness-of-fit test for the model. The null hypothesis is that the vanishing tetrads hold, and a significant test means that at least one of the tetrads does not vanish as expected leading to the rejection of the model.

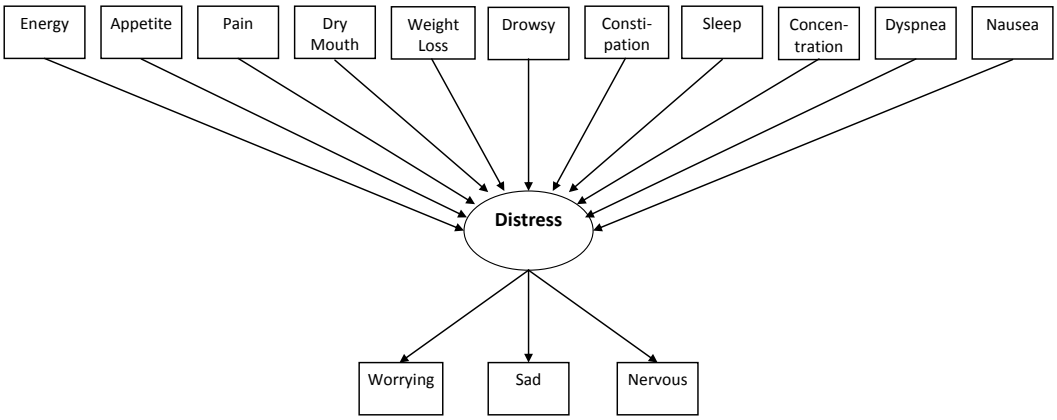
The full list of vanishing tetrads implied by a model is likely to include

redundant tetrads. For example, any two of the three tetrads in Equation 6 implies the third so this model has two non-redundant tetrads. For the simultaneous test redundant tetrads must be eliminated. This means that the set that is retained for testing is not unique and the results of the test may vary based on which tetrads are eliminated. Therefore Bollen and Ting (1993) and Hipp and Bollen (2003) suggest conducting sensitivity analyses.

The advantage of the vanishing tetrad test is that it can be used to compare fit for models that may not be identified or nested in the traditional likelihood ratio sense but may be nested with respect to their tetrads. If the vanishing tetrads of one model is a subset of the vanishing tetrads for another model then they are nested and can be compared statistically. The procedure is similar to the likelihood ratio test in that the difference in the χ^2 test statistic is distributed as χ^2 with degrees of freedom equaling the difference in the number non-redundant vanishing tetrads. This method gets around the problem that formative scales are not identified because a purely formative model has no vanishing tetrads which means it is always nested. The test of a reflectively measured model is equivalent to testing the model against a formatively measured model. If the reflectively measured model is not rejected, Bollen and Ting (1993) recommend further testing of the model using traditional SEM software.

I compared two models. The first model is the bifactor model that was used in the prior analyses. This model has a general distress primary factor and two item-group secondary factors representing physical and psychological symptoms. In light of the literature discussed above with respect to the appropriate treatment of

Figure 5. Hypothesized mixed indicator model.



items in quality of life scales, and the application of Bollen's (1989) mental experiment, I hypothesized a mixed indicator unidimensional MIMIC model, where the physical items are formative and psychological items are reflective. This model is shown in Figure 5. The physical items were kept intact as a group to be consistent with the bifactor model representation that was shown to have good fit to the data. It should be noted that the mixed indicator model is identified, but the two models are not nested in the traditional sense, so the models could not be compared statistically using MPlus' DIFFTEST. I used the research data set for this analysis because too many observations would be lost to listwise deletion due to missing data in the palliative care sample. Listwise deletion occurs with respect to the formative items that become covariates in the model.

Since identification of the non-redundant vanishing tetrads can be tedious with a large number of items, I used a SAS macro developed by Hipp, Bauer, and Bollen (2005) to carry out the tetrad analysis and extend the method to categorical

indicators. Input to the macro includes the polychoric correlation matrix, the asymptotic polychoric correlation matrix, and the model implied covariance matrix or matrices if two models are being compared. PRELIS and LISREL were used to obtain the input matrices. LISREL was selected over Mplus because of the availability of the option to increase the precision in the output listing, which is recommended by the macro's authors, especially for the model implied covariance matrix. As a sensitivity check, the macro was run specifying 15 replications, where each replication represents a random selection of non-redundant vanishing tetrads. Holm's (1979) step-down method was applied to the Sidak (1967) procedure was used to adjust p values for multiple testing.

Results of the Confirmatory Tetrad Analysis

The hypothesized mixed indicator model was fit using Mplus and LISREL, and the fit indices indicated good fit to the data (CFI= 0.996, TLI= 0.982, RMSEA= 0.029). The results of the vanishing tetrad analysis are shown in Table 12. The table has three sets of columns that correspond to the three tests that are carried out when the macro is set up to run a nested test: The simultaneous vanishing tetrad χ^2 statistic for each model and the nested χ^2 difference test. As previously mentioned, the vanishing tetrad test for individual models can be interpreted as a goodness-of-fit test because it is testing whether the observed vanishing tetrads are consistent with the model implied vanishing tetrads. The results for the bifactor model show that it has 20 non-redundant vanishing tetrads and none of the 15 replications were significant so the bifactor model is not rejected. The mixed indicator model has 22 non-redundant vanishing tetrads, and 2 of the 15

replications were significant, but a multiple testing adjustment was not not applied, so there is not sufficient evidence to reject this model either.

The last set of columns show the result of the nested test which compares the two models using a χ^2 difference test. The macro confirmed that the two models are tetrad nested, and the mixed indicator model actually has 2 more non-redundant vanishing tetrads than the bifactor model. A significant test indicates better fit in the model with fewer vanishing tetrads. Based on unadjusted p-values, the results were mixed, with 8 significant tests. After applying the step-down Sidak adjustment for multiple testing, none of the results were significant, providing support for the mixed indicator model.

Table 12

Results of the Confirmatory Tetrad Analysis

	Bifactor Model			Mixed Indicator Model			Nested Tetrad Test			
Replication	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	Adjusted[a] *p*
1	24.33	20	0.229	29.03	22	0.144	4.70	2	0.095	0.379
2	19.59	20	0.484	27.05	22	0.209	7.47	2	0.024	0.199
3	24.66	20	0.215	27.93	22	0.178	3.27	2	0.195	0.379
4	23.86	20	0.249	29.19	22	0.139	5.33	2	0.070	0.371
5	24.39	20	0.226	32.67	22	0.067	8.28	2	0.016	0.176
6	25.67	20	0.177	36.18	22	0.029	10.51	2	0.005	0.072
7	24.05	20	0.240	28.16	22	0.170	4.11	2	0.128	0.379
8	25.63	20	0.178	34.86	22	0.040	9.23	2	0.010	0.131
9	21.35	20	0.377	29.43	22	0.133	8.08	2	0.018	0.181
10	23.45	20	0.267	25.92	22	0.255	2.47	2	0.291	0.379
11	22.98	20	0.290	31.79	22	0.081	8.81	2	0.012	0.145
12	22.55	20	0.311	28.06	22	0.174	5.51	2	0.064	0.371
13	24.48	20	0.222	29.26	22	0.138	4.78	2	0.091	0.379
14	23.28	20	0.275	30.85	22	0.099	7.57	2	0.023	0.199
15	23.12	20	0.283	30.78	22	0.101	7.66	2	0.022	0.199

[a]Adjusted for multiple testing using Holm's (1979) step-down method applied to the Sidak (1967) procedure.

Summary and Discussion of the Directionality Analysis

In this section, I used tetrad analysis to test the hypothesis that the physical items of the CMSAS are more appropriately treated as formative measures. Tetrad analysis is an empirical approach that can be used to test global fit, where fit is defined as the degree to which the data estimated vanishing tetrads are consistent with the model implied vanishing tetrads. The approach can be used to compare models in the same way that the likelihood ratio test is used to compare models. The advantage is that models may be tetrad-nested in situations where the likelihood ratio test can not be used, such as when the models are not nested.

Confirmatory tetrad analysis was used to test both the previously fit bifactor model and the hypothesized unidimensional mixed model that treated the physical symptoms as formative indicators. The individual models were not rejected for poor fit. The comparison of the two models yielded inconclusive results with a slight edge to the bifactor model.

When using confirmatory factor analysis, the model χ^2 test was always significant, even though the other goodness-of-fit measures indicated good fit. This was attributed to the sensitivity of the test to sample size. The tetrad χ^2 test is not sensitive to sample size, and the fact that the test for the bifactor model was not significant was consistent with the global fit indices (CFI, TLI, and RMSEA). It therefore argues for using both approaches when possible.

In this analysis, I only tested the hypothesized mixed indicator model and did not attempt to improve or change this model - though it can be argued that some of the physical symptoms might behave more like psychological items. This

was a confirmatory analysis, not exploratory, and the purpose which the results supported, was to determine if a mixed indicator model was plausible. Furthermore, directionality was investigated after the fact for a scale where items had been dropped on the basis of reflective measurement.

I also did not attempt to evaluate the mixed indicator model. For example, in the mixed indicator model, the effects of the formative items on the reflective items are completely mediated by the latent variable, which implies proportionality constraints on the coefficients (Hayduk, 1996). It should also be noted that issues such as model identification, assessing reliability and measurement error, and construct validity, have cast doubt on the use of formative measurement (see for example, Edwards, 2011).

Tetrad analysis is attractive because it provides an empirical approach for distinguishing between reflective and formative measurement. But the method has not gained widespread use nor has it been implemented in any statistical package. Other approaches for assessing directionality, and for analyzing and estimating parameters, should be explored. MPlus and LISREL can be used to estimate MIMIC models. Other options to be considered are explanatory IRT models (De Boeck & Wilson, 2004), a network approach as described by Cramer, Waldorp, van der Maas, and Borsboom (2010), and partial least squares (PLS; Wold, 1975). The network approach is especially promising for symptoms because it does not require specifying items as reflective or formative and it allows for items to both types at the same time, which is perhaps the most realistic representation.

Guided by the literature (for example, Bollen & Lennox, 1991; Fayers &

Hand, 1997; Streiner, 2003, Zumbo, 2007), especially with respect to quality of life measurement, the application of Bollen's (1989) *mental experiment*, and based on a variety of evidence culminating in the tetrad analysis, I conclude that it is more appropriate to treat many, if not all of the physical items as formative. In other words, physical symptoms are the source of the distress, not the other way around. Edwards and Bagozzi (2005) note that tetrad analysis only allows for tentative conclusions because imperfect reflective measures may not conform to expected patterns, and formative measures may display covariances that are indicative of reflective measures. With respect to symptoms, responses to the items could be reflecting common disease and treatment processes (Fayers & Hand, 1997). Additional evidence also emerged in conjunction with the other analyses which was consistent with the literature. Residual covariation was found among pairs of physical items, and most of these same items were also identified as non-invariant.

I conclude that tetrad analysis may be a useful method for evaluating models, either as a supplement to traditional methods or in situations where models may not be identified or nested. However, determination of reflective or formative measurement should be made *a priori* during scale development based on an examination of the causal mechanisms, and empirical methods should be used for confirmation.

Discussion

Validity Evaluation (Evaluation of the Interpretative Argument)

In this first section of the Discussion, I return to the interpretative argument that was developed earlier and carry out the second step of the argument-based approach which is the validity evaluation. The interpretative argument identified a list of claims or assumptions (see Table 2) upon which the inferences made with respect to using the scale to measure program outcomes depend. The plausibility of each claim is assessed based on the existence and strength of the evidence, some of which was developed in the analyses described earlier. As noted, supporting evidence can be empirical or analytical, and preferred evidence may differ from what is available or feasible. After examining each claim, the interpretative argument as whole is evaluated.

1. The scoring rule is appropriate with respect to equal weighting of symptoms and response options. The assumption that all symptoms contribute equally to symptom distress is generally considered to be reasonable when the items are reflective measures of the construct (Fayers & Hand, 1997), but may not be appropriate for formative items. Therefore, the plausibility of this claim is somewhat dependent on the plausibility of symptoms as reflective measures.

It is not clear if the developers of the original scale and all subsequent versions had explicitly considered directionality of the symptoms as indicators of distress. However, they were treated as reflective by virtue of the fact that the authors' reported Cronbach's alpha, a measure of internal consistency, as a measure of reliability in the published validation studies (Chang et al., 2000; Chang et al.,

2004; Portenoy et al., 1994). Items that do not display this relationship (in other words, salient loadings, inter-item correlations, or item-scale correlations) are theoretically dropped from the scale as part of the development process.

To examine directionality, a mixed indicator model was hypothesized and tested empirically. The analysis showed that while a mixed indicator model is plausible and provided good fit to the data, the all-reflective model could not be rejected either. Although I conclude that the physical items should be treated as formative, I did not find evidence of a need to change the scoring to incorporate unequal item weights.

For the set of response options to be appropriate, there should be adequate endorsement of each option (Embretson & Reise, 2000). If a patient responds affirmatively that a physical symptom is present, the first response option to the question of how much the symptom bothers him or her is "Not at all." The distribution of responses for each item was examined and this option was rarely endorsed. The item response curves showed that this option is almost identical to responding that the symptom is not present. Together they suggest that this option be eliminated. But as I have repeatedly noted, response options have implications for scoring, respondent burden and missing data, and the sensitivity of the scale to detect change, so all three must be taken into account when considering changes. Response options are discussed again below.

2. The observed score is free of bias or inaccuracies introduced by either the clinician or the patient. The scenarios and the potential consequences related to this claim that were detailed earlier were based on

anecdotes, informal observations, and educated guesses about how the palliative care clinicians perform their responsibilities. In the context of measuring improvement in symptom distress, it was noted that completion rates were worse for reassessments, suggesting that clinicians may only be revisiting the symptoms that were present in the initial assessment. If this is the case, then the patient's score at reassessment may actually be worse than the initial score. Therefore, there is sufficient reason to challenge the plausibility of this claim. Evidence for this claim would have to come from an audit-type study, which would be difficult to design and carry out without placing additional burden on patients.

Circumstantial evidence suggesting bias or inaccuracies might be gleaned from the comparison of symptom prevalence between the research and palliative care samples (Table 5). Given that the prevalence rates were higher in the research sample for almost all the symptoms, the conclusion might be that the research sample was sicker than the palliative care sample, but assessment rates in the research sample were nearly 100%. One hypothesis is that the research sample was not sicker, and missing data in the palliative care sample is understating the true level of symptom distress. Another hypothesis is that the scale does not work the same for volunteer subjects as it does for clinical subjects. But, the invariance study suggested that for the most part, the scale does work the same way in both populations.

We rely on clinician reports for evidence to support the assumption that patient responses are not biased. One bias that has been reported is the tendency for veterans to understate physical symptoms, especially pain. However, this bias

exists program-wide, and would not have an impact on comparisons within the program; it would limit the ability to compare outcomes to non-VA programs.

3. The scale is relatively free of measurement error. This claim pertains to random errors that are introduced by irrelevant variation in the measurement procedure. In an operational environment, especially one that spans multiple facilities, and where the scale is also used to aid clinical decision-making, the likely source of random error is inconsistent administration of the scale. Weak support for this claim is provided by the fact that the program uses electronic templates to document palliative care consultations, including responses to the scale. The template encourages but does not ensure consistency. Training has been provided, and is reinforced to a certain extent, through feedback on process measures. Anecdotal evidence is that clinicians have adopted their own approach to administering the scale.

Therefore, measurement error was assessed using confirmatory factor analysis with the rationale that a well fitting model consistent with the expected factor structure would provide evidence that the variable conditions did not substantially undermine measurement precision. As part of the psychometric analysis, I found that the posited bifactor model provided good fit to a subset of the palliative care sample providing support for the generalization inference. The sample was restricted to patients with cancer as their primary disease, and included only the patient's initial assessment. This was done in order to minimize other sources of variation which were addressed in conjunction with other assumptions. The analysis was also restricted by necessity to the binary component of the response, and should

be revisited, if and when the data are more complete.

4. The scale generalizes across contexts. The prior claim was with respect to random error; this claim is about systematic variability in scores attributable to different contexts. As noted, the context for the use of the CMSAS in palliative care differs from the development context two significant ways: it is a clinical operations environment rather than a research environment, and some palliative care patients have diseases other than cancer. The generalization claim was examined by carrying out a series of invariance studies.

In the first invariance study, multiple group CFA was used to assess measurement equivalence between the study sample used to develop the CMSAS, and a comparable sample from palliative care operations. This sample was restricted to initial assessments of cancer patients. The bifactor model fit reasonably well in both samples. The results showed partial measurement invariance for the scale. Six items had invariant loadings, thresholds, and residual variances while eight items were non-invariant. The differences were minor enough to conclude that measurement invariance generally held between the operations and research contexts.

Since the scale was developed using a cancer population, I examined invariance with respect to disease (cancer and non-cancer). The non-cancer population was not large enough to employ multiple group CFA, so MIMIC modeling was used, with a covariate indicating whether the patient was in the cancer or non-cancer disease group. Three items were identified as having DIF: "Pain," "Weight Loss," and "Shortness of Breath." Based on the available evidence,

the scale appears to generalize across a number of contexts.

5. No construct under-representation or irrelevant variance. The more questionable assumption is with respect to construct under-representation. Scales are created and used precisely because of the need to standardize the measurement procedure. Standardization reduces random variation providing support for the generalization claim but at the same time threatening the extrapolation claim. Structure is imposed with respect to three aspects of the symptom distress construct: (a) the list of symptoms included in the scale, (b) the severity domain, and (c) number and wording of response options.

The CMSAS was developed using a sample of cancer patients, and consists of 14 symptoms. Palliative care patients, especially those with diseases other than cancer, may be experiencing distress due to symptoms that are not part of the scale. Likewise, the scale has specified the language used to describe the severity of a symptom. The original scale, the MSAS, had three sets of response options representing three domains of severity (frequency, severity, and bother/distress). For the CMSAS, it was determined that the appropriate response options should be with respect to bother/distress for the physical symptoms, and frequency for the psychological symptoms. Responses to the questions as they are structured may not provide an accurate characterization of the level of symptom distress.

As far as I know, no studies have been undertaken to investigate possible construct underrepresentation with respect to the non-cancer diseases in a palliative care population. There may also be a need for additional psychological symptoms, especially if a mixed indicator model is adopted, and the psychological items are the

only true reflectors of symptom distress. A glaring omission seems to be depression. A new content validation study should be conducted with clinicians having expertise these diseases. Patient input should also be sought with respect to both the list of symptoms and the response options, using focus groups or interviews that allow patients to list and describe their symptoms in their own words.

Construct irrelevant variance can be investigated using the methods applied with respect to generalization, such as multiple group CFA and MIMIC modeling. Also, to a certain extent, construct irrelevance was investigated by examining model fit. This approach may not be effective if the source of irrelevance occurs in a small portion of the population, and is therefore not detected. Invariance analyses should be carried out with respect to other potential sources of variation, such as palliative care team or facility type (nursing home or acute hospital). Also, as palliative care is initiated earlier in the disease process, measurement invariance should be examined in this regard.

These methods require specification of the suspected source of non-invariance *a priori*. Cohen and Bolt (2005) proposed a strategy of identifying latent classes experiencing differential item functioning (DIF) using a mixture IRT model. Once the group is identified, then characteristics that distinguish this group can be investigated. The likelihood is that the group is defined by a constellation of characteristics that otherwise could not have been found. This analysis has not been done, but is planned for a later time when there is a larger sample size with greater diversity.

6. Scale is sensitive enough to detect clinically meaningful changes and 7. The scale is longitudinally invariant. Neither claim could be evaluated in either the palliative care or research samples. The role of response options with respect to change requires at least two assessments over an appropriate time period. The palliative care sample could not be used to assess these claims because of the extent of missing data. Reassessments were often not attempted, or tended to be less complete than initial assessments. It should also be noted that the median number of days from initial consultation to death was 15. Reassessment rates should also improve when palliative care is initiated sooner. The research sample could not be used because subjects were only assessed once.

The completion statistics also seem to indicate that patients are more capable of stating whether a symptom is present, but then have difficulty indicating a severity level. It is not clear if the issue is the number of response options, or that for patients with serious illness, indicating severity level is too great a burden, regardless of the number or wording of response options. On the other hand, there is some evidence supporting the claim that the response options have the sensitivity to register changes in symptom distress. This evidence is the fact that analyses conducted elsewhere showed improvements in mean levels of distress for individual symptoms upon reassessment.

8. Improvement can be attributed to palliative care. This claim relates to the validity of causal inferences in the absence of random assignment and a control group (Cook & Campbell, 1979). Essentially the issues mirror those that are often encountered in evaluation research. One recommended approach is to rule out

Table 13

The Results of Evaluating the Claims of the Interpretative Argument

Claims
1. The scoring rule is appropriate with respect to: Equal weighting of symptoms ✓ The number of response options ?
2. The observed score is free of bias or inaccuracies introduced by either the clinician or the patient ?
3. The scale is relatively free of measurement error ✓
4. The scale generalizes across contexts ✓
5. No construct under-representation or irrelevance ✗
6. The scale is sensitive enough to detect clinically meaningful changes –
7. The scale longitudinally invariant –
8. Improvement can be attributed to palliative care ✓

Note. A ✓indicates that claim is generally supported; a ? indicates that claim is questionable; a ✗ indicates that claim is not supported; a – indicates that claim could not be evaluated

all other alternative explanations for the outcome. At this early stage in the practice of palliative care, the patient is likely to be receiving only palliative care, and therefore the claim is generally tenable. However, as palliative care matures, it will become more common for patients to be receiving palliative care alongside curative care. Eventually this claim will need to be revisited, as it will become less plausible.

Summary of the validity evaluation. Table 13 is a scorecard that indicates the plausibility of each claim in the interpretative argument. To summarize, the evidence shows that the responses to the scale in an operations setting fit the hypothesized factor model relatively well, suggesting that despite the irregularities introduced in a large, diverse clinical environment, measurement using the scale is consistent. Furthermore, the measurement properties of the scale were generally preserved from the more structured research environment in which the

scale was developed. Likewise, while some non-invariant items were identified, the scale retained the same basic measurement properties in a non-cancer population.

I also concluded that the physical items should be treated as formative indicators. However, this finding did not directly affect the validity of the use of the scale to measure program outcomes. But, this new formulation puts more emphasis on the psychological symptoms, where I believe there is construct underrepresentation.

Unfortunately, the degree of missing data due to incomplete assessments prevented a more thorough assessment of the scale, and undermines the use of the scale to measure program outcomes. I could only model the presence/absence component of the responses, which prevented a complete assessment of scoring, the response options, longitudinal invariance, and sensitivity. Missing data is certainly attributed in part to patient limitations, but the research sample statistics show that very ill patients are capable of completing the scale. Therefore, I believe the issue boils down to the fact that the clinicians simply fail to see the scale as a scale, and instead view it as a list of stand alone items. This problem is rooted in the fact that the scale is used to guide clinical care. Any attempts to force completion of the scale for program evaluation purposes is likely to distort the measurement process further.

General Comments on the Argument-Based Approach

In this dissertation, I demonstrated how the argument-based approach to validity could be applied to the validation of a PRO scale, especially in situations where an existing scale is being used for a new purpose or in a new context. The

important thing to note is that it may not be possible to obtain all the needed supporting evidence. In the current situation, I was fortunate to have access to the development sample that was used when the palliative care data was not sufficient. But there were limitations to this data, and it could not be used to assess longitudinal invariance. Often, it will not be feasible because of time or budget constraints to conduct validity studies in advance of using the scale. This situation probably exemplifies the norm, where scales are put into use first and validity is assessed on the back end, if at all. The value of the argument-based approach is that it first examines what claims and assumptions are being made, without regard to what evidence exists or is needed, allowing for a realistic assessment of the validity of the proposed use. The following are a few additional comments and observations on the argument-based approach to validity.

Role of scale developers and scale users. The distinction between the interpretative argument and the validity evaluation was made because the interpretative argument implies an advocacy or subjective role, while the validity argument implies an objective or critical role (Kane, 2006). It can be useful to think of these two roles as roughly corresponding to the roles of developers and users of PRO scales. Traditionally, developers are responsible for conducting validity studies, but they cannot anticipate all the possible uses, populations, and conditions under which the scale will be used. Furthermore, scales are often altered in some way to fit a particular situation, such as eliminating items, changing the wording of items, or the number or wording of response options. PROMIS provides a good example where this explication of roles comes into play. The arduous task of

developing, reviewing, calibrating, and testing items has been organized, centralized, and coordinated resulting in an efficient and thorough process. However, this does not absolve the users of scales from all aspects of validity, especially those aspects that are unique to the particular application and use.

Some validity theorists make a point of distinguishing between validity and validation, maintaining that validity is a property of the test or scale. However, they still still view issues related to the interpretation or use of test scores as part of the validation process. For example, Borsboom, Mellenbergh, and Van Heerden (2004) suggest the use of an umbrella term such as *overall quality* to encompass the important properties such as reliability, predictive adequacy, or absence of bias, that in their opinion, a claim of validity does not imply. In Zumbo's (2009) framework, validity is one of four elements in his "integrative cognitive judgment of validity and validation process" (p. 69), along with utility, psychometrics, and social consequences. This view of validation can be seen as another and perhaps clearer way of defining the roles of the scale developer and the scale user. The scale developer may be responsible for validity, but the user is responsible for measurement quality.

Use of PROs in clinical operations. Caveats have been issued for the use of PROs for clinical operations, mostly with respect to the responsiveness of the instruments (Hays & Hadorn, 1992). But, PRO scales will increasingly be used in this way for a number of reasons. First, there is an ever-increasing call for measuring clinical outcomes (Donaldson, 2008) and for documenting care, especially in areas where clinical outcomes are not readily defined in terms of concrete, clinical

measures such as survival. There is also an increased call for incorporating patient perspectives as outcomes in health care and medical research, in addition to traditional clinical measures (Acquantdro et al., 2003). Hahn et al. (2007) call for formal evaluation of quality of life in routine physical examinations. Therefore, validity in these contexts must be addressed.

Validation of PROs used in research studies. Many of the assumptions that were questionable in the present study may be easier to defend in the context of a research study. For example, evidence for consistent and complete administration may simply be the existence of a research protocol and limiting administration to a very small number of study personnel, who receive very specific instructions and training in accordance with the protocol. Training can include observing appropriate administration, practicing under supervision, and assessing inter-rater reliability. The assumptions are also likely to be different for respondents who are voluntary subjects in research studies and respondents who are patients in clinical programs.

Use of a PRO scale as an outcome measure in a research study is often a one-time use. In these situations, researchers have to rely on existing evidence because it will be unlikely that there will be time or budget to carry out validity studies in advance. However, if researchers perform validity-related analyses after data are collected and include the results in their publications, the body of validity evidence will be enhanced in many ways. Most importantly, it will not consist solely of the evidence provided under the artificial conditions under which the scale was developed.

Current approaches to missing data. Missing data arises when clinicians do not fully administer the scale, or patients lack the stamina or refuse to complete the scale. This problem is well recognized in PRO research, and it is commonly dealt with by (a) making the scoring rule a mean and not a sum and (b) declaring up front some cutoff for the allowable number of missing items and still have the responses "count" (for example, Fayers & Machin, 2007, pp. 355-384). The problem with both solutions is that the assumption is being made that the level of distress for the missing items is the same as mean level of distress for the non-missing items. This approach may be especially problematic if missingness is related to the level of distress for a symptom. It becomes even more of a problem if the construct is formatively measured, since each item provides an independent contribution.

Traditional types of validity. The extrapolation inference probably corresponds the most closely to the traditional validity taxonomy of content, criterion, and construct validity. The difference is that the type of validity evidence should be dictated by the the nature of the extrapolation. For example, correlation with a criterion is appropriate when making an inference about the relationship of test scores to non-test behavior, such as SAT scores and success in college. Otherwise, relationship to a criterion should only be used if the criterion is a gold standard, which rarely exists for the constructs being measured by PROs (Zumbo, 2009). The use of correlations for construct validity is what Cronbach (1988) referred to as a weak program of construct validity, whereas a strong program provides an explanation (Zumbo, 2009).

Issues that were not addressed. One validity-related issue that was not addressed in this dissertation was mode of administration, because the mode is the same throughout the program. The scale is administered by clinicians. However, the scale can and has been self administered. This mode may someday be adopted if and when palliative care becomes more integrated with primary and outpatient care. It would be interesting to see if measurement invariance is maintained in this new context. Likewise, computer administered and computer adaptive applications need to be assessed for invariance, especially in the case of CATs, where subjects are potentially administered a different set of items.

General Discussion

Response options and format. The conditional response format is an artifact of the original MSAS scale, where patients rated symptoms that were present on three dimensions (severity, frequency, and distress), as an attempt to obtain a more complete picture of the impact of symptoms. Based on the degree and nature of missing data, the suggestion might be made to eliminate the conditional structure altogether and replace it with a set of response options where the lowest level corresponds to the absence of the symptom. But, the current approach can be viewed in two ways: (a) the two-part question takes longer to administer, jeopardizing the ability to complete the scale or (b) if patients truly can respond to a yes or no question but not a Likert-type question, then the conditional format might yield more information than if the conditional aspect were eliminated (in other words, a partially complete response versus a totally incomplete response). Furthermore, the responses may be skewed toward absence because patients may

respond "no" if the symptom is not present but provide no response if the symptom is present, but a Likert format is a problem for them. A three-point format, such as "not present," "present with minimal distress," or "present with a lot of distress" may be a good compromise.

Alternative approaches to assessment. The original genesis of this project was to explore ways to further shorten the scale using computer adaptive testing or tailored short forms in order to facilitate fully completed scales. But, two issues quickly emerged. The first was the recognition that the items were not really exchangeable, and the second was that if anything, the scale might need to be lengthened to incorporate more psychological symptoms.

Palliative care is a still maturing field and there has been a subtle shift in focus. Symptom relief, especially pain, is still the primary goal of palliative care, but there has been greater recognition that suffering encompasses something broader which is reflected in the move toward *whole person assessment* (for example, Lacey & Sanderson, 2010). According to the practice guidelines, palliative care should be delivered by an interdisciplinary team consisting of a physician, an advanced practice nurse, a social worker, a psychologist, and a chaplain, so the concept of the whole person approach is not new. But, it was not fully operationalized until we learned from surveys of deceased patients' families how important psychosocial and spiritual support is to these patients and their families.

This shift is not inconsistent with the findings of this dissertation. Indications throughout the analyses pointed towards treating the physical items as formative, while the psychological items consistently behaved as good reflective

measures. Under a mixed indicator formulation, especially if more psychological items are included in the scale, the construct being measured may change. Perhaps, symptom distress would no longer be the correct label and instead, it is a more generic and general distress associated with life-limiting illness consistent with the *whole person* view of suffering.

In this new approach to assessment, physicians are still going to query the patient about their symptoms as part of their clinical care, but perhaps for this purpose, a simple checklist or three-point scale as previously suggested suffices. For symptoms with established efficacious treatments such as pain, the PROMIS item bank could be used to obtain a separate, but more complete assessment of the symptom to guide care.

One of my observations was that the clinicians did not see the scale as a scale. In some ways this new formulation may feel more like a scale than a checklist. Assessment completion may also be improved by treating the distress construct not as a continuous variable, but as a categorical variable or latent class (Collins & Lanza, 2009). The advantage of this approach, is that it is more consistent with how clinicians view the world and provides information to the clinicians in a way that may be more useful to them (Rindskopf & Rindskopf, 1986). Mitchell (2009) contends that it is invalid to treat a construct as measurable if it does not possess an additive structure. An argument can be made that the construct of symptom distress and PRO constructs in general do not meet this definition, so a latent class approach may be a more valid way of characterizing distress.

Data integrity. The focus has been on the problem of missing data and how it threatens the validity of the use of the scale to measure outcomes. This problem can be characterized in a more general way as data integrity. When scales are used for multiple purposes, it can create a conflict of interest when the purposes are potentially at odds with each other. The scale is used by clinicians as a tool to guide care, but how they use the scale and record results could be influenced by the knowledge that the scale is being used to demonstrate the value of the program or judge their performance relative to other programs or palliative care teams. The situation is not unlike the attempt to use value-added models (McCaffrey, Koretz, & Hamilton, 2003) to rate teachers rather than its intended use which was to provide information to teachers to identity student needs.

Contributions made by this dissertation. The current situation is that discussions of validity and validation are at a theoretical level, but the practice of validity and validation for PROs is at a very simple level: perform a few correlations and calculate Cronbach's alpha. Gorin (2007) observed that after attempting to explain the *holy trinity*, *unified framework*, or *argument-based approaches*, her students would be hard pressed to evaluate validity. In this dissertation, I attempted to validate the use of a symptom assessment scale in palliative care to measure program outcomes, using an argument-based approach to validity. The intent was to translate the theoretical description into an actual application and demonstration of the method. The result is what I believe to be the first application of the argument-based approach to validity to the use of a PRO scale .

Another goal of this dissertation was to make the point that while the basics

of validity and validation may be the same across applications, there are things unique to health measurement that make it different. I showed how the argument-based approach is especially useful for identifying the threats to validity that are unique to PROs. In particular, rather than thinking in terms of types of validity where these threats can be missed, developing the interpretative argument provides a broader perspective, where these issues are more likely to come to light.

This dissertation focused on validation from the perspective of the user of scales. Researchers are much more likely, especially in the age of PROMIS, to be adopting and adapting existing scales rather than developing new scales. I made explicit the roles and responsibilities of scale users, which were reinforced throughout the dissertation.

I believe measurement invariance studies are underutilized in general and especially, with respect to validation. In this dissertation, two methods for assessing measurement equivalence were demonstrated in the context of validity. Furthermore, the invariance studies were carried out with respect to a bifactor model. I was able to compare the measurement properties of the scale in a palliative care operations environment to the research environment in which the scale was developed because I had access to the development data sample. Therefore, I call for making the data used to calibrate the PROMIS item banks available to the public for similar studies.

This dissertation included the first application of tetrad analysis to a PRO scale outside the demonstration articles by Bollen and his colleagues. Vanishing tetrads were used to empirically assess whether some of the items of the scale should be treated as formative measures of the distress construct. Appropriate treatment of

items is an issue that researchers are either not aware of or ignore, but has consequences for the development and validation of scales.

Limitations of the study. A limitation of the analyses carried out in this dissertation is that the subjects were almost exclusively male veterans. Many of the analyses would have to be redone in order to confirm the findings in a non-VA population. Perhaps the biggest limitation was the extent of missing data, especially the lack of response to the severity component when the symptom was present. As a result, many of the claims of the interpretive argument could not be evaluated at all, or as thoroughly as I intended. At the same time, the nature of the missing data provided insight into the scale and its response options. The insights are speculative and would need to be confirmed in future research. Another limitation was that because of other planned research, I was not able to obtain feedback from the clinicians regarding the scale, how they administered it, or possible explanations for missing responses. Likewise, I was not able to obtain information on response processes. Input from clinicians and patients will be a critical part of future efforts to improve the measurement of symptom distress.

Appendix

The Condensed Memorial Symptom Assessment Scale

Symptom	Present	Not at all	A little bit	Some what	Quite a bit	Very much
Lack of Energy	Y N	0.8	1.6	2.4	3.2	4.0
Lack of Appetite	Y N	0.8	1.6	2.4	3.2	4.0
Pain	Y N	0.8	1.6	2.4	3.2	4.0
Dry Mouth	Y N	0.8	1.6	2.4	3.2	4.0
Weight Loss	Y N	0.8	1.6	2.4	3.2	4.0
Feeling Drowsy	Y N	0.8	1.6	2.4	3.2	4.0
Shortness of Breath	Y N	0.8	1.6	2.4	3.2	4.0
Constipation	Y N	0.8	1.6	2.4	3.2	4.0
Difficulty Sleeping	Y N	0.8	1.6	2.4	3.2	4.0
Difficulty Concentrating	Y N	0.8	1.6	2.4	3.2	4.0
Nausea	Y N	0.8	1.6	2.4	3.2	4.0

How frequently did these symptoms occur in the last week?

Symptom	Present	Rarely	Occasionally	Frequently	Almost Constantly
Worrying	Y N	1	2	3	4
Feeling Sad	Y N	1	2	3	4
Feeling Nervous	Y N	1	2	3	4

Symptom Distress Score = mean of the 14 items
Physical Distress Subscale = mean of the 11 physical items
Psychological Distress Subscale = mean of the 3 psychological items

References

Acquadro, C., Berzon, R., Dubois, D., Leidy, N. K., Marquis, P., Revicki, D., & Rothman, M. (2003). Incorporating the patient's perspective into drug development and communication: An ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group Meeting at the Food and Drug Administration, February 16, 2001. *Value in Health, 6*(5), 522-531.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Education Research Association.

Asparouhov, T., & Muthén, B. (2006, May 26). *Robust chi square difference testing with mean and variance adjusted test statistics.* [Mplus Web Notes No. 10] retrieved from `http://www.statmodel.com/download/webnotes/webnote10.pdf`

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data.* [Mplus Technical Appendix] retrieved from `http://www.statmodel.com/download/GstrucMissingRevision.pdf`

Bagozzi, R. P. (1982). The role of measurement in theory construction and hypothesis testing: Toward a holistic model. In C. Fornell (Ed.), *A second generation of multivariate analysis* (Vol. 2, pp. 5-23). New York: Praeger.

Bartholomew, D., & Knott, M. (1999). *Latent variable models and factor analysis.* New York: Oxford University Press.

Basch, E. (2010). The missing voice of patients in drug-safety reporting. *New England Journal of Medicine, 362*(10), 865-869.

Bentler, P.M. (1990), Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-46.

Bentler, P.M., & Bonnet, D.C. (1980), Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Boehmer, S., & Luszczynska, A. (2006). Two kinds of items in quality of life instruments: Indicator and causal variables in the EORTC QLQ-C30. *Quality of Life Research, 15*(1), 131-141.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (1990). Outlier screening and a distribution-free test for vanishing tetrads. *Sociological Methods & Research, 19*(1), 80-92.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314.

Bollen, K. A., & Ting, K.-F. (1993). Confirmatory tetrad analysis. In P. Marsden (Ed.), *Sociological methodology 1993* (pp. 147-1750). Washington, DC: American Sociological Association.

Bollen, K. A., & Ting, K.-F. (2000). A tetrad test for causal indicators. *Psychological Methods, 5*(1), 3-22.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: The Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Bucic, T., & Gudergan, S. P. (2004). *Formative versus reflective measurement: Implications for explaining innovation in marketing partnerships.* Paper presented at the Australian and New Zealand Marketing Academy, Wellington, New Zealand.

Byrne, B. M., Shavelson, R. J., & Muthn, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Cella, D., Chang, C. H., Wright, B. D., Von Roenn, J. H., & Skeel, R. T. (2005). Defining higher order dimensions of self-reported health: Further evidence for a two-dimensional structure. *Evaluation and the Health Professions, 28*(2), 122-141.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology, 63*(11), 1179-1194.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., ... Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care, 45*(5), S3-S11.

Chang, C.-H., & Reeve, B. (2005). Item response theory and its applications to patient-reported outcomes measurement. *Evaluation and the Health Professions, 28*(3), 264-282.

Chang, V., Hwang, S., Feuerman, M., Kasimis, B., & Thaler, H. (2000). The memorial symptom assessment scale short form (MSAS-SF). *Cancer, 89*(5), 1162-1171.

Chang, V., Hwang, S., Kasimis, B., & Thaler, H. (2004). Shorter symptom assessment instruments: The condensed Memorial Symptom Assessment Scale (CMSAS). *Cancer investigation, 22*(4), 526-536.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement. 42*, 133-48.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: John Wiley & Sons Inc.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33*(2-3), 137-150.

Crocker, L., & Algina, J. (1986). *Introduction to modern and classical test theory*. Orlando, FL: Holt Rinehart & Winston.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168.

Donaldson, G. (2008). Patient-reported outcomes and the mandate of measurement. *Quality of Life Research, 17*, 1303-1313.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*, 370-388.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174.

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

Fan, X., Thompson, B., & Wang, L. (1999), Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes, *Structural Equation Modeling, 6*(1), 56-83.

Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6*(2), 139-150.

Fayers, P. M., & Machin, D. (2007). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes*. Chichester: Wiley.

Fleishman, J. A. (2004). *Using MIMIC models to assess the influence of differential item functioning*. Paper presented at the Advances in Health Outcomes Measurement: Exploring the Current State and the Future of Item Response Theory, Item Banks, and Computer-Adaptive Testing, Bethesda, MD. Retrieved from `http://outcomes.cancer.gov/conference/irt/fleishman.pdf`

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing research, 19*(4), 440-452.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., & Hays, R. D. (2007). What Is sufficient evidence for the reliability and validity of patient reported outcome measures? *Value in Health, 10*, (Supplement 2), S94-S105.

Gibbons, R. D., Immekus, J. C., & Bock, R. D. (2007). *Didactic workbook: The added value of multidimensional IRT models (Final Report for National Cancer Institute Contract No. 2005-05828-00-00)*. Retrieved from `www.uic.edu/labs/biostat/articles/NCI\_Didactic\_Workbook.pdf`

Glymour, C, Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure*. Orlando, FL: Academic Press.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8), 456-462.

Hahn, E. A., Cella, D., Chassany, O., Fairclough, D. L., Wong, G. Y., & Hays, R. D. (2007). Precision of health-related quality-of-life data compared with other clinical measures. *Mayo Clinic Proceedings, 82*(10), 1244-1254.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. In H. L. Costner (Ed.), *Sociological methodology 1971* (pp. 81-117). San Francisco: Jossey-Bass.

Hayduk, L. A. (1996). *LISREL: Issues, debates, and strategies*. Baltimore, MD: Johns Hopkins University Press.

Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research, 1*(1), pp. 73-75.

Hipp, J. R., Bauer, D. J., & Bollen, K. A. (2005). Conducting tetrad tests of model fit and contrasts of tetrad-nested models: A new SAS macro. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(1), 76-93.

Hipp, J. R., & Bollen, K. A. (2003). Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociological Methodology, 33*(1), 267-305.

Holm, S. (1979), A simple sequentially rejective Bonferroni test procedure. *Scandinavian Journal of Statistics, 6*, 65 - 70.

Holzinger K. J., & Swineford F. 1937. The bifactor method. *Psychometrika, 2*, 41-54

Hu, L.T., & Bentler, P.M. (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling, 6*(1), 1-55.

Jha, A. K., Perlin, J. B., Kizer, K. W., & Dudley, R. A. (2003). Effect of the transformation of the Veterans Affairs health care system on the quality of care. *New England Journal of Medicine, 348*, 2218-2227.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G. (2004). *On chi-squares for the independence model and fit Measures in LISREL*. [LISREL technical document] retrieved from `http://www.ssicentral.com/lisrel/techdocs/ftb.pdf`

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Karnofsky, D. A. & Burchenal, J. H. (1949). The clinical evaluation of chemotherapeutic agents in cancer. In: Macleod CM (Ed.), *Evaluation of chemotherapeutic agents*. New York: Columbia University Press, 1949:191-205.

Lacey, J., & Sanderson, C. (2010). The oncologist's role in care of the dying cancer patient. *The Cancer Journal, 16*(5), 532-541.

LISREL (Version 8.8). [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114*(3), 533-541.

Mazor, K. M., Clauser, B. E., Field, T., Yood, R. A., & Gurwitz, J. H. (2002). A demonstration of the impact of response bias on the results of patient satisfaction surveys. *Health Services Research, 37*(5), 1403-1417.

McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability* [Rand Corporation monograph]. Retrieved from `http://www.rand.org/pubs/monographs/2004/RAND\_MG158.sum.pdf`

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11), S69-S77.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3-62.

Millsap, R. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473.

Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Mitchell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 111-134). Charlotte, NC: Information Age Publishing, Inc.

Mplus (Version 6). [Computer software]. Los Angeles, CA: Muthen & Muthen.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer, & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum Associates.

Muthén, B., du Toit, S.H.C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes* [technical report]. Retrieved from `http://gseis.ucla.edu/faculty/muthen/articles/Article\_075.pdf`

Muthén, L. K. (2010, December 15). Re: Model fit index WRMR [Mplus Discussion]. Retrieved from `http://www.statmodel.com/discussion/messages/9/5096.html?1292455350`

Muthén, L. K., & Muthén, B. O. (2009). *Mplus short courses topic 2 handout* [Course notes]. Retrieved from `http://www.statmodel.com/download/Topic\%202-v14.pdf`

National Consensus Project for Quality Palliative Care (2009). *Clinical practice guidelines for quality palliative care* (2nd ed.). Retrieved from `http://www.nationalconsensusproject.org`

Penrod, J. P., Cortez T., & Luhrs, C. A. (2007). Use of a report card to implement a network-based palliative care program. *Journal of Palliative Medicine, 10*(4), 858-860.

Pentz, M. A., & Chou, C. P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology, 62*(3), 450-462.

Portenoy, R. K., Thaler, H. T., Kornblith, A. B., Lepore, J., Friedlander-Klar, H., Kiyasu, E., ... Scher, H. (1994). The Memorial Symptom Assessment Scale:

an instrument for the evaluation of symptom prevalence, characteristics, and distress. *European Journal of Cancer, 30A*, 1326-36.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer Science+Business Media, Inc.

Reise, S., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16* (Supplement 1), 19-31.

Rindskopf, D., & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine, 5*(1), 21-27.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Sidak, Z. (1967), Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*, 626-633.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*, 201-292.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Sprague, L (2004). *Veterans Health Care: Balancing Resources and Responsibilities.* Washington D.C.: The George Washington University.

Steiger, J. H. & Lind, J. (1980). *Statistically-based tests for the number of common factors.* Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*(3), 217-222.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Teresi, J. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(11), S152-S170.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test validity*. Hillsdale: Lawrence Erlbaum.

Toulmin, S. (1969). *The uses of argument*, Cambridge, England: Cambridge University Press.

Tucker, L., & Lewis, C. (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1):110.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Veterans Health Administration. (2003). *VHA directive 2003-008: Palliative care consult teams (PCCT)*. Washington, D.C.: Veterans Health Administration.

Weissman, D. E., Morrison, R. S., & Meier, D. E. (2010). Center to Advance Palliative Care Palliative Care clinical care and customer satisfaction metrics consensus recommendations. *Journal of Palliative Medicine, 13*(2), 179-184.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324): American Psychological Association.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), *Quantitative sociology: International perspectives on mathematical and statistical modeling* (pp. 307-357). NewYork: Academic Press.

Wools, S. (2008). Evaluation of validity and validation. Paper presented at the 9th Annual AEA-Europe Conference, Hisar, Bulgaria.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (unpublished doctoral

dissertation). Retrieved from
http://www.statmodel.com/download/Yudissertation.pdf

Zumbo, B. D. (2005). Structural equation modeling and test validation. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1951-1958). Chichester, UK: John Wiley & Sons Ltd.

Zumbo, B.D. (2007). Validity: Foundational issues and statistical methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of statistics* (Vol. 26, Psychometrics, pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65-82). Charlotte, NC: Information Age Publishing, Inc.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In David Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.