2015

Clinical Significance of Response Shift in a Spine Interventional Clinical Trial

Robin Carlson
Walden University

Walden University

College of Management and Technology

This is to certify that the doctoral dissertation by

Robin Carlson

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Branford McAllister, Committee Chairperson,
Applied Management and Decision Sciences Faculty

Dr. Robert Kilmer, Committee Member,
Applied Management and Decision Sciences Faculty

Dr. Jean Gordon, University Reviewer
Applied Management and Decision Sciences Faculty

Chief Academic Officer
Eric Riedel, Ph.D.

Walden University
2015

Abstract

Clinical Significance of Response Shift in a Spine Interventional Clinical Trial

by

Robin S. Carlson

MBA, University of Phoenix, 1995

BSEE, United States Air Force Academy, 1983

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Applied Management and Decision Sciences

Walden University

February 2015

Abstract

The effectiveness of treatments for degenerative spine conditions, where the primary symptom is back pain, is typically determined using patient-reported quality of life (QoL) measures. However, patients may adjust their internal standards when scoring QoL based on factors other than their health. This response shift phenomenon could confound the interpretation of study data and impact effectiveness conclusions. In the current study, response shift was examined using structural equation modeling (SEM) and previously collected clinical trial data comparing 2 minimally invasive medical devices in lumbar spinal stenosis patients through 1 year postintervention. In subject QoL results, reprioritization shift between 3 months and 12 months that could confound standard analysis was identified. Treatment group did not influence response shift identified at 12 months. SEM provided an effective and practical tool for clinical investigators to assess response shift in available clinical study data. As response shift could lead to invalid conclusions when QoL measures are analyzed, clinical investigators should include response shift assessment in the design of clinical trials. This research into how response shift phenomenon can impact clinical trial results improves the ability of clinical investigators to interpret clinical trial data, potentially preventing erroneous conclusions. This research may also assist researchers and government regulators in the identification and reimbursement of beneficial, cost-effective medical treatments for patients worldwide. For clinical research designers, this study demonstrates a practical application of response shift assessment.

Clinical Significance of Response Shift in a Spine Interventional Clinical Trial

by

Robin S. Carlson

MBA, University of Phoenix, 1995

BSEE, United States Air Force Academy, 1983

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Applied Management and Decision Sciences

Walden University

February 2015

Dedication

This dissertation is dedicated to my husband, Rick, and my parents, Claire and Doug. Thank you for all your support and encouragement on this scholarly journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to the Study

Health care reform, with its emphasis on evidence-based medicine, has focused increased attention on the findings of clinical trials. Study conclusions influence multiple aspects of health care including not just physicians selecting best treatment options, but also regulators determining marketing approvals, and insurers making reimbursement decisions. When objective biological or physiological markers are not practical or available, clinical trials often rely on patient-reported outcomes (Hamidou, Dabakuyo, & Bonnetain, 2011). These measures, captured through quality of life (QoL) and function questionnaires, can introduce new challenges to data interpretation. Response shift phenomenon, the adaptation over time in the way an individual perceives and scores his or her health quality based on life events, can confound the comparison of longitudinal scores and mask true change in trial data (Donaldson, 2005; Ring, Hofer, Heuston, Harris, & O'Boyle, 2005).

Current researchers of response shift phenomenon have identified response shift in a wide variety of conditions including cancer, stroke, and orthopedic pain (Copay et al., 2010; Oort, Visser, & Sprangers, 2005; Mayo, Scott, Dendukuri, Ahmed, & Wood-Dauphinee, 2008), included theory and definitions (Razmjou, Schwartz, Yee, & Finkelstein, 2009; Sprangers & Schwartz, 1999), and outlined a number of methodologies for response shift identification (Ahmed, Mayo, Wood-Dauphinee, Hanley, & Cohen, 2005; McPhail & Haines, 2010b; Schwartz et al., 2011). However, the clinical significance and practical assessment of response shift in randomized clinical trial data has not been sufficiently investigated (Barclay-Goddard, Lix, Tate, Weinberg, & Mayo,

2009). To provide clinical investigators with an enhanced ability to interpret clinical trial data, research on the relative effect of response shift in comparative clinical trial results is needed.

This chapter contains an outline of a research study of response shift phenomenon in a randomized clinical trial that compared two interventions for lumbar spinal stenosis, a degenerative spine condition. This summary includes an overview of the response shift phenomenon body of knowledge, the research problem statement and purpose, and study assumptions and limitations. More detailed study information including a literature review and detailed study methodology can be found in Chapters 2 and 3.

Background

Clinical trials often use patient-reported outcomes, such as QoL measures, to compare the efficacy and value of medical treatments (Houweling, 2010; Kvam, Wisløff, & Fayers, 2010). However, when patient-reported outcome measures were included in clinical trials, researchers identified paradoxical and illogical findings, such as ill patients reporting the same QoL as healthy individuals (Li & Rapkin, 2009; Wilson, 1999). Researchers have identified that response shift phenomenon, when individuals adapt the way they score their health-related QoL based on factors other than treatment, can complicate the interpretation of clinical trial results (Bernhard, Hürny, Maibach, Herrmann, & Laffer, 1999; Hamidou et al., 2011; Osborne, Hawkins, & Sprangers, 2006; Schwartz & Finkelstein, 2009). Specifically, this adaptation in the way an individual understands and rates his or her well-being has been identified as a confounding factor in

longitudinal studies of medical interventions (Donaldson, 2005; McPhail & Haines, 2010a; Razmjou et al., 2009; Schwartz et al., 2006).

In health care and QoL research, response shift has been defined as a change in the meaning of self-reported outcome measures as a result of recalibration, reprioritization, or reconceptualization (Sprangers & Schwartz, 1999, p. 1508). Recalibration is the adaptation of an individual's internal measurement standards. Reprioritization occurs when the values of a respondent change and reconceptualization is when a subject reinterprets a QoL concept that is used in the construct. As a natural coping process in individuals, response shift can be either the goal of a behavioral intervention, that is, improved adaptation to a chronic disease or disability; or a confounding factor in medical research, that is, when patient-reported outcome measures are compared pre- and postintervention (Schwartz, Andresen, Nosek, Krahn, & RRTC Expert Panel on Health Status Measurement, 2007; Wilson, 1999).

Health care researchers have investigated response shift in a wide variety of diseases and conditions. They have identified this phenomenon as an important behavioral aspect of health care delivery for rehabilitation, geriatrics, and palliative care (Osborne et al., 2006; Yardley & Dibb, 2007). Response shift has also been explored and identified in specific disease conditions including cancer (Bernhard et al., 1999; Hamidou et al., 2011; King-Kallimanis et al., 2012; King-Kallimanis, Oort, Visser, & Sprangers, 2009; Kvam et al., 2010), multiple sclerosis (King-Kallimanis, Oort, Nolte, Schwartz, & Sprangers, 2011; Schwartz et al., 2011), dental treatment (Ring et al., 2005), and orthopedic conditions (Haro, Maekawa, & Hamada, 2008; Razmjou, 2009). The effect of

response shift on study data has ranged from small to moderate but has been shown to have the ability to change clinical conclusions (Ring et al., 2005; Schwartz & Sprangers, 1999; Schwartz et al., 2006).

A variety of methods have been implemented to identify and quantify response shift—design approaches including the then-test, individualized methods, preference-based methods, and statistical approaches (Schwartz & Sprangers, 1999). Because statistical methods allow simultaneous testing of multiple response shift hypotheses and require only the study QoL instruments with no additional patient input, researchers have used these techniques to examine available clinical trial data. The advanced technique of structural equation modeling (SEM) has been successfully used to investigate response shift phenomenon. Based on covariance multivariable regression, SEM has the ability to explore all three components of response shift individually—reconceptualization, reprioritization, and recalibration (Oort, 2005a). The SEM approach has also been applied to address multiple timepoints and exogenous factors (Kline, 2011; King-Kallimanis, Oort, & Garst, 2010; Oort et al., 2005). In health care, SEM has been used to explore response shift in stroke, multiple sclerosis, and cancer (Barclay-Goddard, Lix, et al., 2009; King-Kallimanis et al., 2012; King-Kallimanis et al., 2011, Oort et al, 2005)

This review of response shift phenomenon has highlighted that clinical investigators typically do not evaluate response shift when assessing trial results and do not understand the potential clinical significance of this phenomenon. While there has been noteworthy research into this phenomenon, it is been addressed primarily from a QoL perspective and not in a way that translates this information into current clinical trial

design for randomized clinical trials or spine interventions. Therefore, further examination of response shift with an emphasis on clinical significance and measurement methodologies is needed to help guide comparative medical research and interpretation of clinical trial data.

Problem Statement

In order to improve care and decrease costs, health care reform has increased reliance on evidence-based medicine. For degenerative spine conditions where back pain is the primary symptom, effectiveness is measured by patient-reported health care QoL. However, the internal standards patients use to assess their QoL adapt over time and as a result of their disease. While the scholarly literature has documented this phenomenon in spine conditions (Copay et al., 2010), the clinical significance in randomized medical device clinical trials has not been investigated. Specifically, there have been no studies exploring the potential impact of response shift phenomenon in comparative spine intervention studies when the decision endpoint is 1 year after the intervention.

Purpose of Study

The purpose of this study was to explore the impact on clinical trial conclusions of response shift in lumbar spinal stenosis patients at 1 year postintervention. I evaluated the relationship between the latent variables, physical QoL (PQoL) and mental QoL (MQoL); and observed QoL variables from the Zurich Claudication Questionnaire (ZCQ), Short Form General Health Survey (SF-12), Oswestry Disability Index (ODI) and the pain Visual Analog Scale (VAS), in a structural equation model to assess response shift. Using secondary data from baseline, 3 months, and 1 year, I explored the clinical

significance of response shift by comparing this phenomenon between 3 months and 1 year and between two treatment groups.

Nature of Study

I used a quantitative longitudinal confirmatory modeling research design and QoL data collected at three timepoints to support the investigation of the clinical significance of response shift. The data used for this investigation were a subset of the data collected as part of a randomized clinical trial comparing two lumbar spine medical devices. To evaluate response shift, I performed confirmatory SEM techniques as a secondary analysis on previously collected experimental data. By comparing response shift between 3 and 12 months and determining if there was a difference in the phenomenon identified between these two points in time, I gained insights into the importance of considering response shift in clinical data interpretation. In addition, investigating differences in response shift at different timepoints increased my understanding of the role time plays in this phenomenon. By evaluating response shift between treatment groups at 12 months, I explored if a randomized clinical trial design changed the importance of considering response shift in clinical studies. I selected this research study design because the study population and modeling technique were aligned with the research requirements and the methodology could be applied to existing data as a secondary analysis. By implementing SEM techniques, I also assessed the practicality and usefulness of using this methodology for response shift investigation.

To effectively investigate response shift, a population where response shift can be expected was required. Spine treatment research has documented inconsistencies

consistent with response shift phenomenon (Copay et al., 2010; Schwartz & Finkelstein, 2009). Patients who receive a lumbar spinal stenosis surgical intervention also meet the criteria for a health state catalyst as defined by Sprangers and Schwartz (1999). A change in health state is expected because the intervention is designed to relieve vertebral pressure and reduce symptoms in patients currently experiencing debilitating back and leg pain. Finally, as lumbar spinal stenosis is a progressive disease that interventions do not fully resolve, the treatment effect is often partial and not total. Incomplete resolution has also been identified as a factor in populations where responses shift is more likely to be found (Finkelstein, Razmjou, & Schwartz, 2009).

This research also required an evaluation technique that supported the integration of multiple observed variables and was sufficiently sensitive to identify the phenomenon of interest. The statistical analytical approach of SEM met this criterion. When compared to other available statistical techniques, SEM can address both measurement and conceptual issues in a single model (Oort et al., 2009), can identify all three components of response shift (Schwartz et al., 2011), and can address multiple follow-up timepoints (Hamidou et al., 2011). This method does not require response shift specific data collection and has been demonstrated in the literature (Barclay-Goddard, Lix, et al., 2009; King-Kallimanis et al., 2011; Oort et al., 2005). Because few current comparative clinical trials have addressed response shift directly, the ability to apply this methodology to previously collected data increased the value of the research to on-going clinical trials. For these reasons, a SEM technique using data from an on-going spine intervention study was identified to support this investigation.

Instead of independent and dependent variables, in structural equation models the variables are termed latent and observed variables. The latent variables, factors not directly measured, for this research included PQoL and MQoL. Observed variables included the scores from QoL measures including the ZCQ, SF-12, ODI, and the pain VAS. The exogenous variable of primary interest, a factor external to the model, was treatment group, either the control or the investigational intervention. Additional exogenous variables that were included in the exploratory analysis included age at time of surgery, body mass index (BMI), gender, and number of levels treated (1 or 2).

The observed variables were provided from an on-going prospective, randomized, controlled, multicenter trial titled Investigating Superion™ Interspinous Spacer in Spinal Stenosis (ISISS). The ISISS study enrolled approximately 470 patients who (a) were at least 45 years of age, (b) experienced moderate symptoms of neurogenic claudication secondary to lumbar spinal stenosis, and (c) had documented stenosis at one or two contiguous levels from L1 to L5 (VertiFlex, 2013). The ISISS trial completed enrollment in 2012 and the data were submitted to the U.S. Food and Drug Administration (FDA) as part of a premarket application. In the ISISS study, subjects who met all eligibility criteria were randomized to either the investigational or a control device (1:1 ratio) and had a medical device implanted. At each follow-up (6 weeks, 3 months, 6 months, 12 months, 18 months, and 24 months), study sites collected subject adverse event information, radiographic images, a neurological assessment, and subject function and QoL questionnaires. For this secondary research into response shift, only QoL data through 12 months were included in the dataset and the specific treatment group

(investigational or control) was blinded (coded as A or B) to safeguard the results of the comparative performance of the two medical devices.

I used the SEM technique as presented by Oort (2005b) and King-Kallimanis et al. (2010) to test for response shift in the ISISS prospectively collected data. Patient-reported QoL and function data from subjects enrolled in the trial, that had a device implanted, and were followed for 12 months were used in the three-step process. First, a measurement model was established using SEM best practices (Kline, 2011). Then, I tested the model for invariance across the measurement occasions of baseline, 3, and 12 months to identify response shift. Finally, I tested the model for invariance with respect to the exogenous variables. The response shift results from the SEM analysis were then used to address the research questions and hypotheses. Details of the research questions and hypotheses are found in Chapter 3.

Research Questions

This study was designed to answer the following research questions:

1. Do treated back pain patients experience a difference in response shift between baseline and 3 months and between baseline and 12 months postintervention?
2. Does response shift phenomenon influence the clinical comparison of patient-reported outcomes between baseline and 12 months in a randomized clinical trial for a spine intervention?

Hypothesis 1

$H_0$1: Response shift at 12 months is not different from response shift at 3 months.

$$RS_{12} = RS_3$$

H_a1: Response shift at 12 months is different from response shift at 3 months.

$$RS_{12} \neq RS_3$$

Hypothesis 2

$H_0$2: Response shift found in the patient-reported outcome results for treatment group A at 12 months is not different from the response shift in treatment group B at 12 months.

$$RS_A = RS_B$$

H_a2: Response shift found in the patient-reported outcome results for treatment group A at 12 months is different from the response shift in treatment group B at 12 months.

$$RS_A \neq RS_B$$

Theoretical Foundation

This research was based on two foundational concepts—response shift theory and SEM. Response shift, a psychological adaptation to situations, was first identified in the 1970s in research in management science and educational training (Schwartz, Sprangers, & Fayers, 2005). Modern investigation of response shift was facilitated in the 1990s by the updated theory proposed by Sprangers and Schwartz (1999) that translated research into QoL measures associated with the evaluation of disease progression and assessment of medical treatments. The earliest descriptions of response shift focused on internal

changes in an individual's standards of measurement (Howard & Dailey, 1979) and separating true change from changes in internal standards and meaning (Golembiewski, Billingsley, & Yeager, 1976). Sprangers and Schwartz expanded these concepts to provide a working definition of response shift with a goal of supporting the development of reliable and valid measures for assessing changes in QoL measures. Response shift addresses the paradoxical and counter-intuitive research findings in the QoL literature including consistent discrepancies between clinical measures and patient-reported QoL. Grounded in control theory and self-regulating systems, the theoretical model included a dynamic feedback loop where an individual's behavioral processes used to handle life events (mechanisms) worked with the individual's characteristics (antecedents) in a way that resulted in a QoL result that differed from the expectations based on objective criteria (response shift). Catalysts in this theory included changes in the patient's health status, either positive or negative. Antecedents are a person's stable characteristics such as gender and personality, and mechanisms refer to cognitive processes that the patient uses to adjust to life changes. Finally, QoL is the measurement construct that scores the patient's feelings about his or her life and response shift is a person's change in perspective when evaluating his or her QoL. QoL researchers have identified recalibration, reprioritization, and reconceptualization as components of response shift (Sprangers & Schwartz, 1999). The theoretical foundation of response shift is explored in more detail in Chapter 2.

Modeling provides the methodological foundation for this research. SEM is a family of multivariate analysis procedures that focus on means, variances, and

covariances to explore relationships between observed and latent (unobserved) variables. Confirmatory factor analysis, a SEM model, is used to determine if collected data fit a theory-based measurement model. Based on covariance, the procedure is designed to accomplish two goals: (a) to understand the patterns of covariance and (b) to use the researcher-specified model to explain as much of the variation as possible (Kline, 2011). SEM is implemented by hypothesizing a causal model, depicting the model as a path diagram, and testing the model using empirical data. This technique incorporates measurement error as latent variables and allows variables to be indicated by multiple measures. The combination of structural path models, factor analysis, and covariance analysis enables SEM to address the complex interactions typical in social science research questions.

The variables that can be included in a structural equation model support the complexity of social science research. Variables may be observed or unobserved (latent) with some of the latent variables representing measurement error (Kline, 2011). The scale of a latent variable is arbitrary and the researcher must set the value in the model. By setting the variance of a latent variable to 1, the scale can be standardized. Alternatively, the variable may take on the scale of one of its indicator variables (Lei & Wu, 2007). These conventions allow for simplification because when fixed in either manner, the variables are not estimated from the data. Similarly, because for endogenous variables all effects are included in the model, no unanalyzed associations occur between these variables (Kline, 2011).

When applied to response shift and health care research, SEM strengths include the ability to incorporate multiple variables into the model, the inclusion of measurement error, and the fact that this technique can be applied without additional data collection. Finally, because in the social sciences the magnitude of the effect is often most important and not the specific result of the statistical test, SEM provides better estimates of effect size for observed variables than many other mathematical techniques (Kline, 2011). Additional discussion of SEM and a detailed methodology for this research study are included in Chapter 3.

Definitions

Correlation matrix: Representation of data in a structural equation model for programming and model specification (Kline, 2011, pp. 48-49).

Endogenous variable: A model component (variable) that is influenced by another variable in the model—a dependent variable (Kline, 2011, p. 96).

Exogenous variable: A model component (variable) not influenced by another variable in the model—an independent variable (Kline, 2011, p. 95).

Health-related quality of life (QoL): "A multidimensional concept that usually includes self-report of the way in which physical, emotional, social, or other domains of well-being are affected by a disease or its treatment" (Calvert et al., 2013, p. 815).

Latent variable: A model component (variable) that is not directly measured or observed. Latent variables can also be called factors (Ullman, 2006, p. 36).

Measurement model: A system depiction (model) that relates measurement variables to latent variables (Ullman, 2006, p. 37).

Observed variable: A model component (variable) that is directly observed or measured. Observed variables can also be called indicators, measured variables, or manifest variables (Ullman, 2006, p. 36).

Response shift phenomenon: A change in the meaning of an individual's self-reported outcome measures such as QoL and function. Specifically, "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (scale recalibration, in psychometric terms); (b) a change in the respondent's values (i.e. the importance of component domains constituting the target construct); or (c) a redefinition of the target construct (i.e. reconceptualization)" (Sprangers & Schwartz, 1999, p 1508).

Structural equation modeling (SEM): A set of modeling techniques based on correlation and matrixes with the potential to differentiate between observed and latent variables (Kline, 2011, pp. 7-9). This research focuses on a common SEM technique, confirmatory factor analysis (CFA), that integrates a measurement and a structural model.

Structural model: A system depiction (model) that documents the hypothesized relationships between all the variables of the model (Ullman, 2006, p. 37).

Quality of life (QoL): "The general well-being of a person or society, defined in terms of health and happiness, rather than wealth" (Collins English Dictionary, n.d.).

Scope of Research

Through this research, I explored the potential impact of response shift phenomenon in a comparative spine intervention with a decision endpoint 1 year after the intervention. The primary ISISS study represented the population of patients in the

United States that is seeking minimally invasive treatment for radiographically confirmed moderate lumbar spinal stenosis. Lumbar spinal stenosis is the most common indication for spinal surgery in adults over age 65 with lumbar spinal stenosis in 20% of U.S. adults over age 60 and 80% in those over age 70 (Loguidice, Bini, Shabat, Miller, & Block, 2011). This response shift research, a secondary evaluation, represented the same population as data from all subjects that completed 12-month follow-ups as of February 2012 were included in the analysis.

Delimitations

Data from a single lumbar spinal stenosis intervention study were used to support this research. All subjects were enrolled between June 2008 and Feb 2012 from 32 geographically dispersed sites in the United States (Loguidice et al., 2011; VertiFlex, 2013). Generalization concerning response shift in other intervention studies, other spine conditions, and other diseases were not considered. Additionally, adverse event information, radiographic, and neurological evaluation results were not be correlated with QoL data because it was beyond the scope of the research.

Assumptions

A key assumption of this research was that the primary clinical study data from ISISS were high quality and could be used to support this response shift research. As a subset of previously collected data, the limited dataset has the same quality characteristics as the original research. Because the ISISS study was designed to provide data for an FDA submission for commercialization, it included compliance with good clinical practice and all federal regulations associated with medical device clinical studies

including 21 Code of Federal Regulations (CFR) Parts 50, 54, 56, and 812 (VertiFlex, 2013). The electronic case report form data collection system was in compliance with 21 CFR Part 11. As required by 21 CFR 812 and good clinical practice, quality control techniques, periodic site monitoring, and source data verification were included in the data collection process.

Another assumption was that the methodology used to assess response shift was sufficiently sensitive to the phenomenon. Because sample size and power analysis are not applicable to advanced modeling techniques such as SEM to determine sensitivity, other support was required. A comparative evaluation of three statistical techniques for measuring response shift identified SEM as being the most successful and providing interpretable findings (Schwartz et al., 2011). Best practices for implementing SEM were integrated into the analysis process and are addressed in more detail in Chapter 3.

Limitations

This research was based on secondary analysis of previously collected experimental clinical trial data. As such, this research design did not drive the data collection. Additionally, the data for this research were collected prior to final quality checks so there was a possibility of data entry errors. These concerns have been mitigated by ensuring that the source data were collected in accordance with good clinical practices, included electronic data entry with built-in edit checks, and had been monitored by the sponsor. Additionally, data screening and cleaning procedures were used prior to analysis.

Advanced modeling also introduces limitations to a research design. Consistent with standard modeling techniques, the determination of goodness of fit and respecification of the structural equation model is dependent on the researcher's subjective judgment. The research design minimized this limitation by prespecifying alternative models and ensuring all model adjustments were supported by theoretical justifications (Kline, 2011).

Significance of Study

This research provides clinical investigators with a practical application of a response shift assessment in a clinical trial and adds to the body of knowledge on response shift phenomenon. For clinical investigators, especially spine researchers, the role of response shift in study findings is highlighted and instruments most likely to be impacted by response shift are identified. This increased knowledge on how to improve clinical trial design to enable the separation of true change from patient adaptation will result in clinical investigators making more accurate conclusions and enhance clinical decision-making. This improved clinical evidence would support the approval and reimbursement of effective, cost-effective treatments and interventions.

This evaluation of response shift in a lumbar spinal stenosis population provides QoL researchers with additional data from patients with a chronic condition and addresses the impact of time as a catalyst variable. The demonstration of a practical methodology supported the integration of response shift assessment into clinical trial design. The results supported better understanding of how patients respond to changes in their health state and enable physicians to support improved QoL adjustment for

individuals living with chronic diseases or disabilities. Enhanced understanding of QoL results and a practical methodology could also be translated back to management research where patient-reported QoL information was included in the analysis.

Summary

Response shift phenomenon, changes in a patient's internal standards when assessing health-related QoL over time, may impact clinical trial results and conclusions used to support health care decision-making. This chapter examined response shift, including current theory, methods of evaluation, and the potential impact of response shift on the data interpretation, and presented a research study to explore this phenomenon. I identified the use of mathematical modeling, specifically SEM, as one method to assess and determine the incidence and impact of response shift in clinical data. The research problem, hypotheses, assumptions, potential limitations, and research definition were outlined. A review of the significance of this research to society concluded the chapter. Further explanation and clarification will be found in subsequent chapters.

Response shift may present a significant confounding factor in clinical trials that use patient-reported outcomes as primary endpoints. However, much remains to be characterized about this phenomenon. This research into the impact of response shift in a spine intervention clinical trial provides valuable information to health care and QoL researchers and includes a practical methodology for response shift assessment. To explore the body of knowledge on response shift phenomenon in more depth, Chapter 2 contains a comprehensive literature review of this phenomenon. In Chapter 3, I outline

the detailed methodology and provide the results of the analysis in Chapter 4. Finally, I discuss the findings and provide practice and future research recommendations in Chapter 5.

Chapter 2: Literature Review

To translate new medical research and technologies into effective treatments and enhanced patient care, clinical trials play a pivotal role. However, when patient-reported outcomes serve as the primary study endpoints, confounding factors may complicate or even invalidate the interpretation of study results (Donaldson, 2005; Hamidou et al., 2011). One potential influencing factor is response shift phenomenon—the fact that, over time and based on life events, individuals can change the way they perceive and report their well-being based on factors unrelated to their health state (Sprangers & Schwartz, 1999). To draw effective conclusions from randomized, comparative health care data, researchers must understand the impact and clinical significance of this phenomenon. Studies of degenerative spine treatments, where pain is the primary symptom and effectiveness is measured by patient-reported health care QoL measures, may be impacted by response shift (Don & Carragee, 2008; Schwartz & Finkelstein, 2009). Only by incorporating the assessment of response shift into clinical trial design will clinical investigators and researchers be able to consistently and effectively assess the best, most effective treatment options for patients.

To explain paradoxical and counterintuitive findings between objective measures and patient-reported outcomes, QoL researchers developed response shift theory (McPhail & Haines, 2010a; Schwartz et al., 2006). In health care research, this phenomenon has been identified in a wide variety of conditions including cancer (Oort et al., 2005), stroke (Mayo et al., 2008), and lumbar spinal stenosis (Copay et al., 2010). In separate comprehensive overviews, Schwartz and Sprangers (1999), Ahmed et al. (2005),

McPhail and Haines (2010a), and Schwartz et al. (2011) identified a number of methodologies used to investigate and quantify response shift including design approaches, individualized methods, preference-based methods, and statistical techniques. The mechanisms and processes of response shift have been researched in many ways though the clinical significance to randomized clinical trial data and the effective transfer of assessing this phenomenon have not yet been supported. A promising approach is the statistical method of SEM (King-Kallimanis et al., 2010; Kline, 2011; Oort et al, 2005). Based on the working definition proposed by Schwartz and Sprangers (1999), this review analyzes response shift in health care research. The key aspects associated with this phenomenon include the theory and foundations, response shift in health care and spine research, measurement approaches, and SEM as applied to response shift. This review provides insights to assist clinical investigators, clinical trial designers, and health care professionals to understand the clinical importance of response shift and to effectively integrate the assessment of response shift into trials thereby improving the quality of care for patients.

I conducted a comprehensive literature search to support this review. Using three primary databases, EBSCO Academic Search Premier, ProQuest Complete, and Science Direct, I searched for articles using the terms *response shift, health related QoL, patient-reported outcomes, measurement bias, SEM, statistical analysis, longitudinal data, orthopedics,* and *spine clinical trials* and combined these terms using Boolean operators to better target relevant articles. I gave preference to peer-reviewed articles published between 2008 and 2014; however, I also included earlier foundational research,

specifically from 2005 and 1999. I screened the bibliographies of the research articles to identify additional sources, key words, and relevant concepts and used Zotero, a bibliography management program, to categorize and document articles.

Response Shift Theory

Experiencing and responding to change is part of every individual's life experience. In the past 15 years, physicians and medical researchers have become increasingly interested in the way a person's health state affects his or her QoL perception, especially changes from treatment or disease progression. Sprangers and Schwartz (1999) defined response shift phenomena as the adaptation of a respondent's internal standards, values, and conceptualization of life quality due to changing health or other life events. Used to understand and explain unanticipated or illogical relationships between objective measures and patient-reported health care QoL, the discipline of response shift is still in an early developmental stage. The complexities of response shift phenomenon and researchers' conceptual and operational confusion have created challenges in research. Despite this, the body of knowledge is diverse and has supported multiple theories, a variety of assessment methodologies, and increased understanding of the clinical significance of response shift in clinical trials, especially in the area of QoL assessment.

Origin of Response Shift Theory

Howard and Dailey (1979) first identified response shift bias while documenting the benefits of educational training programs. Envisioned as an instrumentation effect, the researchers argued that because the purpose of a training program was to change a

subject's knowledge of a specific variable, typically an attribute being measured, a successful program would alter the individual's perspective and impact self-evaluations. With a change in a respondent's internal measurement standards and scales, the assumption of a common yardstick for the two assessments would be violated. The individual's adaption could then confound or even invalidate the standard technique of direct comparison of pre- and posttraining self-evaluations. Recognizing the same phenomenon in organizational research, other early theorists classified and defined types of change. Golembiewski et al. (1976) postulated three types of change—alpha, beta, and gamma. Alpha change was identified as true change, the goal of the treatment or intervention. Beta and gamma change were potential bias factors that indicated scale recalibration and construct reconceptualization. Through statistical analysis, the authors concluded that a single concept of change was not appropriate and noted that failure to account for individual adaptation could result in inaccurate data interpretation and erroneous conclusions.

Building upon these foundations, Sprangers and Schwartz (1999) translated response shift into health care research and developed a theoretical model based on control theory. An individual's psychological adaptation to illness was modeled as a continuous feedback system with the goal being the individuals feeling as good as possible about themselves and their life. The model included five primary variables—catalysts, antecedents, mechanisms, response shift, and perceived QoL. Catalysts were the changes in an individual's health state, regardless of source or direction. These changes could be the result of sudden onset of illness, initial diagnosis, disease

progression, or treatment. Antecedents were the built-in, existing characteristics of the individual such as gender, personality, and spiritual identity. Mechanisms were the processes the individual would use to respond to the catalyst including coping, social support, and goal reordering. Response shift was the adaptation of the person's meaning of QoL resulting from changes in internal standards, values, or well-being concepts. Perceived QoL was the final multidimensional construct of the person's well-being that incorporated physical, mental, social, and other aspects of living. Sprangers and Schwartz theorized that when faced with a change in health status (catalyst), an individual will use known mechanisms influenced both directly and indirectly by antecedents to trigger response shift. The divergence of reported QoL from objective criteria-based expectations would therefore be directly impacted by response shift. The model, highlighting the feedback loop within the individual, is presented in Figure 1.

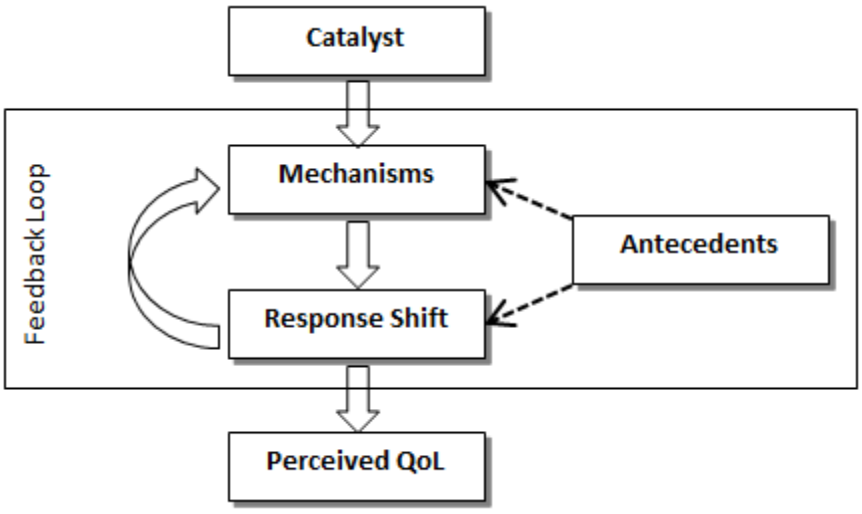


Figure 1. Sprangers and Schwartz response shift feedback loop. Based on description in "Integrating response shift into health-related quality of life research: A theoretical model," by M. Sprangers and C. Schwartz, 1999, *Social Science & Medicine*, *48*, pp. 1507-1515.

Alternative theories of response shift have also been proposed. Carver and Scheier (2000) described a theory for the recalibration component of response shift based on adaptive self-regulation. In the model, when an individual experienced a health state adversity associated with distress that could not be resolved, the individual's well-being reference point would scale back or down. For example, a person who ran competitive 10K races would not expect to compete while they were healing from a broken leg. By reducing expectations and aspirations, the person would again experience both positive and negative feelings around the new set point replacing the mostly negative feelings based on the old reference. Carver and Scheier viewed the change as automatic, outside of voluntary control, and relatively slow. The authors supported the importance of human response shift by arguing that without recalibration, individuals would not be able to overcome life-changing adversity and concluded that human well-being was tied to having purpose in life, and purpose was supported by response shift. Addressing response shift differently, Wilson (1999) proposed investigating how the dimensions of subject well-being including biological measures, symptoms, functioning, and general health perceptions, change relative to each other. In this construct, response shift would be considered an effect that cannot be attributed to the disease or to known mechanisms, therefore a placebo effect. Wilson also identified that response shift was a goal of routine clinical practice since it supported the physician or non-conventional practitioner in allowing the patient to cope with their current health state. Despite the alternative viewpoint, the author validated the importance of response shift in clinical care and

concluded that producing response shift involved understanding the context of the illness, could be impacted by the physician-patient relationship, and could improve patient QoL.

As an early discipline, the conceptual and operational definitions of response shift theory are still being explored. Complementary and competing theories, such as crisis theory and implicit theories including recall bias, impression management, and effort justification have been identified (Mayo et al., 2008; Sprangers & Schwartz, 1999). Researchers have proposed multiple directions for future research to address theory, application, and reporting. Multiple researchers identified a requirement to build a consensus on terminology (Barclay-Goddard, Epstein, & Mayo, 2009, Reeve, 2010; Schwartz et al., 2005). Schwartz et al. (2005) also highlighted that exploring the timing and nature of catalysts and investigating the relationship between the components of response shift would support expanded theory. Integration into health care clinical trials would benefit from research into surrogate markers, additional measurement and adjustment techniques, optimal follow-up timepoint identification, and bias-minimizing study designs (Hamidou et al., 2011). Standard reporting of critical response shift parameters to include data on effect size, sample size and response rates, and details of instrument psychometrics would support the transfer of knowledge both within and across disciplines (Schwartz et al., 2006). Finally, Schwartz et al. (2006) emphasized that researchers should always provide information on how the study results could impact future research and, if applicable, provide clear interpretation guidelines for generalization.

Definition of Response Shift

Sprangers and Schwartz (1999) provided a working definition of response shift that has grounded most health care research into this phenomenon. The authors specified that response shift referred to a change in the meaning of self-reported outcome measures as a result of recalibration, reprioritization, or reconceptualization (p. 1508). Recalibration occurs when an individual changes their internal measurement standards, i.e. the worst pain imaginable reference point gets higher after the patient experiences kidney stones. Reprioritization ensues when an individual changes what they value, i.e. family support and social connections become more important than appearance after chemotherapy results in hair loss. Reconceptualization transpires when an individual changes how they interpret a specific concept, i.e. originally physical function and mental health were primary components of QoL, after disease diagnosis and treatment the level of fatigue experienced better defines a good day. This definition has provided a foundation for the investigation and understanding of response shift in health care.

Response shift researchers have been challenged by the complexity of the phenomenon and potential overlap with other coping phenomena. Schwartz et al. (2005) felt the need to address a very basic question—is response shift an umbrella term for different phenomena that should be studied separately or is it an important new concept for QoL research? While noting further research and theory were required, the authors pointed out that if response shift could explain away all paradoxical or unexpected results from patient-reported data then it would have no meaning. Consequently, the researchers recommended that alternative theories should be explored before labeling a finding as

response shift. In the QoL research community, a wide variety of operational definitions have been associated with response shift even when the same working definition was applied. To support analysis, Schwartz et al. identified six primary areas of dispute concerning the dimensions of response shift. These areas covered how best to define response shift and what uniquely identified the phenomenon. The issues were not mutually exclusive and are outlined in Table 1.

Table 1

Response Shift Perspectives

	Perspective 1	Perspective 2
1	Bias (to be corrected for)	Meaningful change (worth investigation)
2	Measurement characteristic (error in instrument selection)	Subject characteristic
3	Ad hoc explanation (for illogical findings)	Phenomenon (to be studied)
4	Temporary change	Permanent change
5	Result of a catalyst	Result of passage of time
6	Events unrelated to health state may be relevant to response shift	Only health-related events are relevant to response shift

Highlighting that when comparing two QoL scores the goal of QoL researchers was to determine what was change in QoL and what was measurement error, Ubel, Peeters, and Smith (2010) proposed that the term response shift be replaced with something less ambiguous. Arguing that since response shift as currently used lumped together measurement error and true QoL changes, Ubel et al. claimed the use of a single term impeded QoL research by creating conceptual confusion. The researchers also identified logical circularity since the operationalization of the mechanism was the same

as the operationalization of response shift and that there was an implied connotation that response shift was a threat to the validity of self-reported data. Other researchers including Sprangers and Schwartz (2010) and Reeve (2010) agreed that there was conceptual confusion surrounding response shift but disagreed with the primary recommendation of abandoning the term. Sprangers and Schwartz highlighted that abandoning response shift terminology would not resolve the outlined problems as it mixed human adaptation processes and measured outcomes and that new approaches, some already in press, could address areas of concern. Reeve concurred with Sprangers and Schwartz that a change of terms was not the solution but agreed with Ubel et al. that there was confusion surrounding the phenomenon. Reeve's recommendations included focusing on educating researchers on terms and the implications of incorporating QoL change as endpoint variables. Researchers could improve research quality by including in study planning an assessment of (a) the QoL construct, (b) the measurement instrument, (c) research study design, (d) the subjects participating in the study, and (e) the investigator interpretation of the data.

Expansion to Theory

Expansions to response shift theory have focused on the methods individuals use to grade their well-being (the appraisal process) and classifying the measurement and explanation components of response shift. Rapkin and Schwartz (2004) adapted a psychological adjustment process to create a four-step procedure for determining an individual's appraisal process included (a) assessing the frame of reference, (b) identifying a subjective sampling strategy, (c) recognizing standards of comparison, and

(d) understanding the combinatory algorithm. The authors concluded that recognizing the appraisal process was a step in relating patient-reported outcomes to external observer perspectives and was an important aspect of understanding response shift phenomena. In related research by the same authors, Schwartz and Rapkin (2004) created a psychometric model integrating the appraisal process into QoL true score assessment and evaluation of questionnaire properties. To improve research, the authors proposed designing new instruments that included appraisal aspects, using appraisal features to stratify subjects, and including appraisal constructs into clinical trials as moderating variables.

In another theory expansion, Oort et al. (2009) explored and defined the concepts of response shift as either measurement or explanation bias, two separate but related ideas researchers have used to classify response shift. From a measurement perspective, response shift would be a special case of measurement bias. Measurement bias involved the idea that the instrument used, the questionnaire, does not perfectly measure the attribute of interest. Therefore due to this mismatch, changes in the test results were not fully determined by changes in the characteristic of interest. This would identify response shift as a difference in the scores that was not explained by true differences but was also not random, a violation of measurement invariance. From a conceptual perspective, response shift would be a special case of explanation bias. Here there was a distinction between predictors of the construct and other variables that might impact the construct. This perspective identified response shift as variance in the scores that was not explained by the model variables but was also not random, a violation of explanation invariance. Oort et al. identified that the primary difference in these two perspectives would impact

the classification of the type of response shift that had been realized in a specific situation—recalibration, reprioritization, or reconceptualization. The researchers concluded that since both perspectives had meaning and value, response shift measurement methodologies that addressed both aspects would be preferred for response shift investigation, especially if the components of response shift could be investigated both jointly and separately. By investigating and classifying these perspectives, response shift theorists have distinguished different schools of thought concerning response shift, helped to resolve confusion in the research community, and highlighted different research methodologies.

Response Shift in Health Care

Any instrument that captures self-reported data from individuals may be impacted by response shift (Wilson, 1999). And with the importance of evidence driving the integration of QoL assessments into clinical research and paradoxical findings indicating that further exploration was required (Oort et al., 2009; Sprangers & Schwartz, 1999), the health care research community has been motivated to explore this phenomenon in greater detail. Reviewing the role of the QoL research community, use in behavioral research, and disease specific studies will provide important background for response shift investigation.

QoL Research

As treatment and intervention decisions are increasingly guided by evidence-based medicine, physicians and clinical researchers have developed methods to assess how patients perceive their care and well-being. Health-related QoL has been used to

assess the clinical benefit of interventions and recognized as a valid clinical endpoint by the medical community (Barclay-Goddard, Epstein, et al., 2009). In cancer research, the U.S. FDA has recommended that health-related QoL assessments be included in clinical research since 1985 (Hamidou et al., 2011). As patient-reported outcomes have grown to encompass patient symptoms, functioning, perceptions of health status, and overall well-being, the QoL research subspecialty has developed to increase knowledge in this discipline. In order to explain counterintuitive QoL findings, researchers from this discipline introduced response shift and have supported significant research into this phenomenon in health care (Sprangers & Schwartz, 1999; Wilson, 1999). Questions asked by the community (Oort et al., 2009; Schwartz et al., 2005) include

- Is response shift associated with measurement or subject characteristics?
- Should response shift be corrected for or studied?
- Is response shift an ad hoc explanation for counterintuitive findings?
- What initiates response shift?

Schwartz et al. (2005) identified three primary reasons that response shift phenomenon should be measured. First, failure to account for response shift could lead to inaccurate treatment effect conclusions in clinical trials. Second, response shift could impact determination of minimal clinical significant change. This concept used to identify thresholds for meaningful improvement recognized that just because change was measureable did not mean it was useful or valuable to individuals. Finally, measuring response shift would provide insight into the human adaptation process resulting in more ways to support patients in effective chronic diseases and disabilities accommodation.

With a focus on the phenomenon of response shift itself, the body of research has been primarily academic. Empirical and clinical data have been presented to illustrate relevant response shift concepts and not to develop practical methods of incorporating response shift into clinical trial design (Barclay-Goddard, Epstein, et al., 2009; Schwartz et al., 2006).

Behavioral Research

An important aspect of clinical care has been determining how effective interventions have been in modifying the behavior and medical outcomes of patients. This was also an area where response shift was first identified. Health care researchers have therefore explored response shift in a variety of applications including modification of behavior and coping with chronic conditions.

Behavior modification for fall prevention, pulmonary rehabilitation, and cancer survivor intervention has been examined for the presence and impact of response shift. McPhail and Haines (2010b) evaluated response shift in a study of hospitalized patients participating in a fall prevention program. Two QoL questionnaires were completed, one within 3 days of hospital admission and one immediately prior to discharge. This older population, mean age of 73.3 years, had a median length of stay of 38 days (range 20-60). Three methods of measuring change were used to analyze the results—a standard pre-test/post-test comparison, a patient perceived change rating method, and a perceived change adjusted for recall bias. The comparison identified a difference between the methods. With 83.2% of the individuals reporting a clinically significant discrepancy between the direct comparison and perceived scores, the authors concluded inaccurate

representation of change in patient self-reported health states could result in unsupported conclusions. In the evaluation of a pulmonary rehabilitation and self-management program for chronic obstructive pulmonary disease (COPD) patients, Ahmed, Bourbeau, Maltais, and Mansour (2009) identified response shift between baseline and 12 months with one statistical method but not with a second. The unadjusted model underestimated average change in physical health though the effect was small. Ahmed et al. found that even small response shift biases could impact effect size—moving the results from small to moderate or moderate to large. The researchers concluded that response shift should be identified and accounted for before study results could be accurately interpreted.

Schwartz, Feinberg, Jilinskaia, and Applegate (1999) examined response shift in young cancer survivors who participated in a 3-day training program. Using an age matched cohort of healthy subjects, a comparison of pre-test to post-test collected immediately after the training and 3 months later, suggested that there was an immediate gain in QoL with a significant decline at the 3-month follow-up. The analysis supported the hypothesis that the results were influenced by response shift. Adjusting for the shift, the researchers found that the intervention changes the survivor's concept of QoL so that it more closely matched the healthy controls. In behavior modification research, the identification of response shift has been mixed. Most studies indicated some level of response shift, with one study reporting both response shift and no response shift depending on the measurement tool (Ahmed, Bourbeau, Maltais, and Mansour, 2009). While effect size was not specifically measured in these studies, in Schwartz et al., when

response shift was considered, the impact of the intervention was adapted from negative to positive.

The management of chronic conditions is another area where response shift has been researched. Nolte, Elsworth, Sinclair, and Osborne (2009) explored response shift in a group-based chronic disease self-management program population. Specifically, the researchers used measurement invariance tests to compared pre-test to post-test and retrospective rating to post-test data. QoL was measured at the start of the course and at the end of the course, typically 6 weeks. The researchers concluded that response shift was indicated in the retrospective rating to post-test data and not in the pre-test to post-test data. This was contrary to the study hypothesis and to response shift theory. Nolte et al. concluded that using a pre-test to post-test comparison was appropriate for assessing the change in this population over 6 weeks and that response shift need not be considered.

Yardley and Dibb (2007) identified response shift in a subpopulation of subjects with Meniere's disease, a moderately disabling chronic illness, by analyzing QoL data measured 10 months apart. The subpopulation with severe symptoms demonstrated response shift while patients with moderate symptoms did not. An individual's social comparison results predicted greater response shift and longer time since diagnosis, higher self-esteem, and increased optimism were associated with less response shift in this population. Osborne et al. (2006) identified response shift in an arthritis chronic disease management program population and developed a paper-based questionnaire to measure response shift. Using individualized interviews as the gold standard and QoL input from 2 to 10 months post course for 121 participants, the researchers validated the

measurement instrument and identified that response shift occurred in approximately 50% of the replies. Of these respondents, 30% realized during the course they had scored their original QoL too high, negative response shift, and 20% realized they had scored their original QoL too low, positive response shift. Both Yardley and Dibb and Osborne et al. concluded that response shift could confound pre-test to post-test assessments and potentially obscure true treatment effects.

In an investigation of traumatic brain injury, Blair, Wilson, Gouick, and Gentleman (2010) identified that, contrary to expectations, there was no significant difference between current and retrospective judgment of past QoL. Researchers asked subjects directly about their change in QoL. Individual differences in responses existed with some subjects reporting that prior to their injury, QoL was better and some that their QoL was worse. When analyzed based on current disability status, a majority of subjects who were not disabled reported no change in their QoL before and after the injury. On the other hand, patients who were disabled reported changes in their QoL with approximately 60% reporting lower QoL while 40% reported increased QoL. Blair et al. identified response shift as the most likely reason for the change in QoL reported. In the area of disabled patients, Schwartz et al. (2007) identified that most QoL measures equate function with health resulting in lower QoL in patients with disabilities regardless of their individual perception of health.

In chronic diseases, the effect of response shift has been mixed. The phenomenon was identified in three studies that investigated longer timeframes, though in one study only a specific subpopulation demonstrated measureable response shift. In a study that

measured response after 6 weeks, no response shift was identified. Details of effect size were not specifically evaluated or reported for these studies.

Disease and Condition Research

Response shift has also been researched in a variety of specific diseases and conditions including cancer (Bernhard et al., 1999; Hamidou et al., 2011; King-Kallimanis et al., 2009, Kvam et al., 2010), stroke (Ahmed et al., 2005; Barclay-Goddard, Epstein, et al., 2009; Mayo et al., 2008) and multiple sclerosis (Schwartz et al., 2011). While most studies focused on determining the existence of response shift, some also specifically investigated effect size, recalibration, reprioritization, and reconceptualization.

As cancer studies often include QoL instruments, several cancer studies have been used to investigate response shift. In a research overview, Hamidou, Dabakuyo, and Bonnetain (2011) identified ten clinical trials published between 1999 and 2010 that illustrated response shift and covered cancer populations with colon, prostate, breast and multiple myeloma. Methods of treatment including surgery and chemotherapy were highlighted as potential catalysts to the phenomenon. The review examined response shift measurement techniques and found that 80% of the trials incorporated the then-test, a technique that explored response shift by having respondents assess their current and retrospective QoL at each follow-up. In a colon cancer study, Bernhard et al. (1999) identified that patients reframed their perception of QoL but that that side effects of chemotherapy did not induce greater response shift. Patients assessed QoL before surgery, after surgery, and after randomized follow-up therapies using retrospective

ratings. Bernhard et al. found that despite the negative impact of chemotherapy, change in subject perceptions was not significantly impacted by treatment and that adjusting for response shift diluted the treatment effect but amplified overall improvement in most QoL indicators.

King-Kallimanis et al. (2009) identified response shift using a measurement perspective but no response shift in an explanation view in newly diagnosed lung, pancreatic, esophageal, and cervical cancer patients who were scheduled for surgery. Using a QoL questionnaire completed prior to and 3 months after surgery, patients changed the way they conceptualized bodily pain and general health after surgery. The researchers noted that the general health construct, derived from questions that were general in nature, seemed to be vulnerable to bias. Using explanation bias guidelines, optimism and upward comparison variables impacted reported mental well-being, however because the effect was consistent across the measurement timepoints, this bias was not interpreted as response shift. In a multiple myeloma population, Kvam et al. (2010) examined minimally important clinical difference estimates and identified response shift changes in both magnitude and direction. At baseline and 3 months, 239 subjects who self-identified as either improved or deteriorated, completed QoL questionnaires. The results of the two groups were compared and researchers found pain, fatigue, and physical function were retrospectively perceived by deteriorating subjects as higher at baseline—subjects minimized their prior QoL. On the other hand, improving subjects had no significant changes in their responses and the authors concluded that for this group the results were robust with no requirement to account for response shift.

Adjusting the deteriorating subject group for response shift would have increased the resulting change in QoL results between baseline and 3 months by 12 points. These findings were consistent with other research comparing improving and deteriorating subjects and identified that response shift may impact study subpopulations differently. Though the effect may be concentrated in specific subpopulations, response shift has often been identified in cancer studies. Surgery and longer follow-up timepoints seem to be associated with identification of this phenomenon.

Stroke populations have also been investigated for response shift. Mayo et al. (2008) identified response shift in a third of the subjects using data from a longitudinal stroke study of 387 subjects that collected data at 1, 3, 6, and 12 months post-stroke. The study, which compared response shift methodologies, found that 67% of the subjects showed no response shift, 15% negative response shift (expected a higher functional level than immediately post-stroke), and 13% had a positive response shift (adapted to a lower normal functional level). Based on validation against retrospective ratings and simulation analysis, the authors concluded that ignoring response shift could yield an acceptable model over 12 months post-stroke. Because the methodology used was group based, the subgroups that experienced response shift would need to be large and in the same direction to be identified. Because the goal of post-stroke rehabilitation was to provide subjects with the coping tools to regain their wellbeing, Mayo et al. noted that identifying that 80% of the population either experienced no response shift or raised their criteria for QoL could be a negative finding since patients were not adjusting to their new health status and limitations.

Using data from a post-stroke study that investigated the effectiveness of at-home treatments, Ahmed et al. (2005) found conflicting response shift estimates when comparing three response shift methodologies—SEM, retrospective ratings, and individualized tests. Based on the retrospective rating and individualized tests, response shift was identified. The SEM method did not identify response shift. On the other hand, Barclay-Goddard, Lix, et al. (2009) used SEM to identify response shift at both 6 and 12 months using data from stroke patients. Focusing on the mental health components of QoL, data from 677 participants who completed five QoL questionnaires at 1, 3, 6, and 12 months post-stoke was analyzed. The researchers identified that this population demonstrated response shift in one aspect of the scale at 6 months and five aspects of the same scale at 12 months. Barclay-Goddard, Lix et al. concluded that recalibration response shift occurred, however effect size was not estimated. These results indicate that clinical significance and effect size of response shift in stroke populations has been varied.

Response shift has also been assessed in multiple sclerosis and other medical conditions. Schwartz et al. (2011) used outcome data from multiple sclerosis patients to investigate and compare three statistics-based response shift quantification techniques—SEM, latent trajectory analysis, and recursive partitioning analysis. Data from 3,008 subjects in the North American Research Committee on Multiple Sclerosis (NARCOMS) data set were classified as relapsing, stable, and progressive without relapse based on the last 2 years of data and current thinking on disease progression. The authors concluded that the detection of response shift was dependent on the technique used. Researchers

using recursive partitioning identified all three aspects of response shift, SEM researchers identified only recalibration, and those using latent trajectory analysis concluded that 99% of the population did not experience response shift. In general, the research found little evidence of response shift in the multiple sclerosis population. The limited response shift findings were unexpected since this population was selected for this research because as a chronic and progressive neurological disease, it was considered likely to include response shift. Schwartz et al. highlighted that these null findings could reflect a true absence of response shift or could be a limitation in the research design or methods.

Research by Li and Rapkin (2009) supported the existence of response shift in an HIV population. A qualitative evaluation of HIV choices in care study provided insights into response shift through an assessment between baseline and 6 months of individual appraisal characteristics (Rapkin & Schwartz, 2004). The researchers identified nine subgroups of respondents that provided insight into cognitive assessment and response shift. King-Kallimanis et al. (2010) also investigated the measurement of response shift using data from HIV/AIDS patient QoL data. Using questionnaires completed every 6 months for 2 years, the authors identified four examples of measurement bias representing three findings of response shift. However, the researchers could not identify any theoretical justification or explanation for these findings so concluded these findings could represent chance results. King-Kallimanis et al. noted that the data did not include a catalyst event, required by current response shift theory, and that could impact response shift.

In a medical device study, Ring et al. (2005) did identify response shift that impacted treatment effect. In this investigation, the researchers compared implanted versus conventional dentures using baseline and 3-month follow-ups measured by two response shift methodologies. The researchers identified both reconceptualization and reprioritization response shift and found that with traditional analysis no treatment effect was demonstrated while adjusting for response shift identified a significant treatment effect.

A significant number of investigations into a variety of health care conditions have been conducted. Researchers have identified a wide variety of findings with follow-up timing and measurement technique seeming to impact the results. However, while literature provides sufficient evidence to indicate that response shift occurs, no consensus has been reached on the impact of the phenomenon on clinical trials data interpretation.

Spine Research and Response Shift

Spine and orthopedic conditions present a unique population for response shift research because no biologic or radiographic markers have been found to effectively assess the result of spine treatment (Don & Carragee, 2008). Therefore the primary, and often only, goal of treatment is the reduction of pain and improvement of QoL with spine clinical investigators using function and QoL questionnaires as primary research endpoints (Copay et al., 2010). Standard practice has been for investigators to compare patient-reported function and QoL before and after an intervention and provide several years of QoL follow-up to comply with FDA requirements. Based on a clinically relevant threshold value, an intervention was interpreted as successful if the level is met.

Measuring QoL in spine trials may be complicated. First, it is often not possible to blind patients to their treatment. In particular, medical device trials must address the ethics and subject expectations when one arm of the study involves surgery and the other conservative therapy. Additionally, subjects may be provided with new information and strategies for coping with the consequences of their degenerative condition when follow-up covers longer periods of time. Finally, as an elderly population, the comorbidities of aging may also impact perceptions. Adaptation of a patient's QoL reference standards for any of these reasons may complicate the interpretation by eliciting response shift. Researchers have identified paradoxical results that suggest response shift may need to be considered in spine clinical trial interpretation (Anderson & Gerbing, 1988; Copay et al., 2010).

Spine Research

Common in North America, chronic low back pain is becoming an increased burden on the health care system (Don & Carragee, 2008). However, despite its prevalence, and probably as a result of the complexity of the condition, there are no standard clinical practice guidelines for surgical intervention. Researchers have evaluated a wide variety of measurement instruments including condition-specific, disease-specific, and general health-related QoL tools. To identify the most effective instruments, Walsh, Hanscom, Lurie and Weinstein (2003) analyzed ODI, Musculoskeletal Outcomes Data Evaluation and Management System (MODEMS), and short form general health survey (SF-36) results completed at baseline and 3 months. After classifying subjects as improved, worsening, and no change and using receiver operating characteristic (ROC)

curve analysis, the researchers concluded that pain scales have been more responsive than function scales and there was no significant difference between condition-specific and general health scales. Therefore, the general SF-36 may be sufficient for low back pain studies and there was no requirement for condition-specific instruments. Copay et al. (2008) investigated minimum clinically important change in this population and validated minimum detectable change for the ODI as 12.8 points, for the physical component score (a subscale of the SF-36) as 4.9 points, for VAS back pain as 1.2 points and for VAS leg pain as 1.6 points.

Spine researchers have reported inconsistencies in their research. When comparing multiple instruments in the same spine population, Copay et al. (2010) found subjects showed considerable inconsistency with 60% reporting conflicting results between ODI, SF-36, back pain VAS, and leg pain VAS collected at baseline and 1 year. However based on an overall change index, the subjects' opinions on their treatment were strongly correlated and the authors concluded that the inconsistencies did not invalidate the QoL measurements. Copay et al. hypothesized that the threshold of tolerable pain could be translated into meaningful information and noted that the inconsistency between QoL instrument results should be taken into account when outcomes were evaluated and clinical relevance assessed. Haro et al. (2008) compared objective measures and QoL instrument results in spine surgery patients at baseline and 24-months post-surgery. The objective measures correlated with the leg pain VAS but not with the back pain VAS or ODI. The SF-36 indicated significant postoperative improvement, in both physical and mental component summaries. However, the objective measures were significantly lower

than the mental health subscale and significantly higher than the physical health subscale. Noting that VAS scores assessed physical health better than mental health and that ODI was most indicative of a patient's symptoms, the authors concluded that patient subjective assessments were important in the evaluation of treatments that focused on the improvement of QoL.

Researchers and theorists have identified response shift as a potential explanation. Using a case study, Schwartz and Finkelstein (2009) outlined basic response shift concepts and applied them to spine research including an assessment of disease trajectory where the treatment resulted in a partial and not a total cure. In spine conditions, resolving one set of symptoms may unmask other pre-existing symptoms, for example the resolution of leg pain exposes back pain. Overestimation of baseline disability may have no clinical significance when a total cure has been realized; however with only a partial cure bias in the baseline may impact measured treatment effect. Anderson, Carreon, and Glassman (2009) also identified incomplete recovery and progressive degeneration as factors in spine conditions that could impact subject satisfaction and cause them to underestimate the effectiveness of an intervention. Finkelstein et al. (2009) also highlighted the potential impact of a partial cure in commentary on orthopedic surgery results. With one group underestimating their baseline measure of impairment and the other overestimating this same baseline, the authors identified two directions of response shift recalibration. Finally, in a study of total knee replacement, Razmjou et al. (2009) directly investigated the existence, direction, and effect of response shift at 6 months and 12 months postintervention. Using both disease specific and general QoL

instruments, the researchers identified response shift at both timepoints with an increase identified from 6 to 12 months. The change was particularly pronounced in the mental component of the SF-36 QoL results. Traditional analysis did not indicate any improvement in mental state as a result of the treatment at either timepoint. However, the values adjusted for response shift demonstrated a statistically significant difference at 12 months.

ISISS Study

The ISISS study is a U.S. investigational device exemption (IDE) pivotal study that used a prospective, randomized, multicenter trial design to compare the safety and efficacy of a minimally invasive spine implant in the treatment of lumbar spinal stenosis (Loguidice et al., 2011; VertiFlex, 2013). Lumbar spinal stenosis is a progressive degenerative spine condition that is experienced by the patient as low back pain with leg pain and weakness during standing and walking. These symptoms result in impaired mobility, limitations in daily activities, loss of independence, and reduced QoL. As a degenerative disease, lumbar spinal stenosis symptoms can be treated both medically and surgically but cannot be cured. Therefore, the primary goal of treatment is to maximize function and maintain QoL.

Subjects at least 45 years old who had experienced a minimum of 6 months of moderate spinal stenosis symptoms and were unresponsive to conservative care were enrolled into the ISISS study (VertiFlex, 2013). After signing an informed consent, subjects were randomized (1:1) to either the investigational or the control device. The device was implanted and subjects were scheduled for follow-up visits approximately

every 6 months for a period of at least 24 months. At follow-up visits, subjects were assessed for neurological function and adverse events, had radiographic imaging, and completed QoL and function questionnaires including the ZCQ, ODI, SF-12, and VAS for back and leg pain. The primary endpoint was success at 24 month based on clinically significant patient improvement as determined by the ZCQ and if postintervention treatment was required. Secondary endpoints are improvement denoted by satisfaction score, ODI, VAS, and general health status. Data quality was supported by electronic case report forms with built-in error checks and regular site and data monitoring in accordance with good clinical practice and FDA regulations. The ISISS study was approved by Western IRB (central) or local site IRBs and all subjects provided informed consent for their data to be used for research purposes. Western IRB determined that the dataset provided for this analysis met all the criteria for a deidentified dataset and that no additional approval from the subjects was required for this secondary analysis.

Measuring Response Shift

In discipline overview articles, Schwartz and Sprangers (1999) and Barclay-Goddard, Epstein, et al. (2009) identified a wide variety of methods that had been implemented to investigate response shift. The methodologies can be grouped into prospective approaches, where response shift assessment is incorporated into the research design, and statistical approaches, that can be implemented as a secondary analysis. Prospective approaches include retrospective ratings, individualized methods, and preference-based methods. Statistical approaches include ANOVA analysis, growth curves interpretation, latent trajectory analysis, and SEM. An overview and assessment of

these methods will be presented. Due to the early stage of response shift development, not all methods have the same level of implementation or empirical evidence.

Retrospective Ratings

One of the first and most commonly applied design approaches for assessing response shift has been retrospective ratings, commonly called the *then-test* (Barclay-Goddard, Epstein, et al., 2009; Schwartz & Sprangers, 1999). In this method, subjects complete the QoL instrument at baseline (the pre-test) and at follow-up (the post-test). In addition, at follow-up the respondent provides a retrospective assessment of their baseline QoL, the then-test. This method has been based on the concept that since the post-treatment QoL and reassessed baseline QoL would be completed at the same time, the respondent would be using the same QoL internal standards, values, and concepts. A comparison of the post-test and then-test would minimize any bias and indicate unconfounded treatment effect. The difference between the baseline (the pre-test) and the retrospective score (the then-test) indicate the level and direction of response shift, specifically recalibration (Ahmed et al., 2005; Schwartz et al, 2005, Schwartz & Sprangers, 1999). The then-test methodology was initially developed to assess educational training programs but has been adapted to be used in health care (Schwartz & Sprangers, 1999).

Health care researchers in a number of medical conditions have used the then-test to explore response shift including studies of cancer (Hamidou et al., 2011; Jansen, Stiggelbout, Nooij, Noordijk, & Kievit, 2000; Kvam et al., 2010; Schwartz, Feinberg, Jilinskaia, and Applegate,1999; Visser, Oort & Sprangers, 2005), stroke (Ahmed et al.,

2005, chronic disease (Yardley and Dibb, 2007), orthopedics (Razmjou et al., 2009), and dental implants (Ring et al., 2005). The researchers documented that retrospective ratings were relatively easy to administer and to interpret. The reliability of the then-test approach is similar to the original measurement tool so most instruments with strong psychometric properties can be used (Schwartz & Sprangers, 1999). The body of literature supports the value and usefulness of integrating then-tests into response shift research.

Despite extensive usage, researchers have also identified potential concerns with the then-test. As patients may not be able to accurately recall their health from months earlier and so incorrectly report prior health, investigators have been concerned that this technique may identify both recall bias and response shift. In an investigation of patient care facilities, McPhail and Haines (2010b) identified a clinically meaningful change between longitudinal results and patient retrospective results in 83.2% of the subjects. However, when the data was adjusted for recall bias, the change was reduced to 7.9%. Based on a review of studies, Barclay-Goddard, Epstein, et al. (2009) identified that how the individual considers the renewed judgment of prior health can impact the findings. If the respondent recalls their prior health state and answers based on that information, the then-test would be accurate. However, if the respondent internally calculates prior health based on perceived change, as predicted by implicit theories of change, the result may be biased. Adding to the mixed results, Nolte et al. (2009) tested the psychometric performance of the then-test in chronic disease management and concluded that while the

standard pre-test/post-test was robust, the performance of the then-test/post-test could be influenced by implicit theory of change, social desirability, and recall bias.

Other considerations when using the then-test include the requirement for subjects to complete multiple questionnaires and for the additional instruments to be incorporated into initial study design (Barclay-Goddard, Epstein, et al., 2009). Finally, the tested population is also required to meet minimum cognitive and memory ability requirements (Ahmed et al., 2005; Schwartz & Sprangers, 1999). Based on these concerns, Reeve concluded in a 2010 commentary on response shift that QoL researchers agreed that better methodologies than the then-test existed for the assessment of response shift.

Individualized Approaches

Individualized methods are measurement techniques that enable the respondent to define and assess the aspects of QoL that are the most relevant to them by integrating subject feedback with specific QoL factors. Instruments, such as the Schedule for the Evaluation of Individual Quality of Life (SEIQoL) and Patient Generated Index (PGI) are most often formatted as semi-structured interviews where the subject selects their five most relevant QoL domains and scores them (Schwartz & Sprangers, 1999). Subjects may identify the domains on their own or select from a reference list. After treatment or over time, the subjects reaccomplish the assessment. Hamidou et al. (2011) highlighted that using these tools researchers could generate an overall index, document the relative importance of each domain at each timepoint, and identify changes in relevant domains over timepoints.

Used by researchers to evaluate response shift in a range of health care situations including cancer (Hamidou et al., 2011), head injury (Ahmed et al., 2005; Blair, Wilson, Gouick, and Gentleman, 2010), comparison of treatment (Ring et al., 2005), and disease self-management (Osborne et al., 2006), the primary advantage of individualized approaches is the focus on the unique, individualized construct of QoL perception. Each participant can identify the aspects of QoL that are relevant to them at the time of the assessment. By comparing changes in domains and weights, researchers can identify specific differences over time and explore more about the meaning of the response shift (Barclay-Goddard, Epstein, et al., 2009; Schwartz et al., 2005). However, the complexity of the individualized assessments may restrict the populations that can be tested since subjects need to have significant cognitive ability to be able to provide relevant input (Blair et al., 2010). Additional disadvantages of this methodology include converting results into numerical values can be difficult (Barclay-Goddard, Epstein, et al., 2009), effective comparison can be challenging since context is changing (Schwartz & Sprangers, 1999), and the time-consuming interview requirement can make the approach impractical for larger studies and some disease states (Hamidou et al., 2011).

Preference-Based Methods

Related to individualized approaches, preference-based methods of assessing response shift ask respondents to rate the value of specific health states such as the acceptable tradeoffs between longevity and a specific health aspect. This approach includes the Q-Twist method, preference mapping, and the ideal scale approach (Hamidou et al., 2011; Schwartz & Sprangers, 1999). In the ideal scale approach, patients

are asked to report both their current and ideal QoL on the same scale. Hamidou et al. (2011) noted that when repeated at different timepoints, researchers could compare the ideal results and identify recalibration and reprioritization response shift.

Preference-based approaches have been used to evaluate response shift in cancer (Hamidou et al., 2011; Visser et al., 2005) and AIDS/ HIV patients (Schwartz et al., 2005). However, while labeled a promising approach in 1999 (Schwartz & Sprangers), only a small number of researchers have used preference-based techniques to investigate response shift in current health care literature. Ease of implementation and resistance to recall bias are advantages of preference-based approaches (Schwartz et al., 2005). However, difficulties of this methodology include the potential for ceiling effects that may limit the results and interpretation may be complicated since both recalibration and reconceptualization are integrated into a single construct (Hamidou et al., 2011). Implementation also often requires advanced statistical techniques (Schwartz & Sprangers, 1999). This approach can also be time intensive for both participants and clinical investigators.

Statistical Approaches

Statistical approaches are the final general methodological classification used by researchers to evaluate response shift. These techniques use statistical tools to identify trends in research data. Statistical techniques have been as simple as paired t-tests (McPhail & Haines, 2010b) and as advanced as SEM (Barclay-Goddard, Epstein, et al., 2009). Additional statistical techniques have included growth curves, regression trees, and latent trajectory analysis. Growth curves techniques have been used to compute and

compare individual slopes on variables of interest. The flexibility of this method and ability to assess relationships between two or more curves enabled growth curves to support complex problems (Barclay-Goddard, Epstein, et al., 2009; Schwartz & Sprangers, 1999). Regression tree analysis used repeated classification techniques to divide the research population into subpopulations that were more homogenous (Li & Rapkin, 2009). These subgroups enabled researchers to gain increased insight into response shift. Latent trajectory analysis created a predictive General Health model and examined discordances between predicted and observed results (Mayo et al., 2008).

Statistical approaches have been used by researchers to evaluate response shift in a variety of medical conditions including cancer (King-Kallimanis et al., 2009; King-Kallimanis et al., 2012; Oort et al., 2005; Visser et al., 2005), stroke (Ahmed et al., 2005; Barclay-Goddard, Lix, et al., 2009), multiple sclerosis (Schwartz et al., 2011), COPD (Ahmed et al., 2009), and HIV/AIDS (Li & Rapkin, 2009). Being model based has provided statistical approaches with a key benefit since it allowed researchers to test multiple response shift hypotheses at the same time (Schwartz & Sprangers, 1999). Data collection and research design were also simplified as only data from standard QoL instruments already included in the study were required. The ability to perform response shift analysis after the data has been collected, as secondary research, provides another significant benefit since comparative clinical trials that did not originally incorporate response shift assessment may investigate this phenomenon. The disadvantages of statistical approaches include results that may not be easily interpretable for non-statisticians (Schwartz & Sprangers, 1999) and the potential for the components of

response shift to be integrated in a way that makes disentangling the results difficult (Mayo et al., 2008). Additionally, statistical techniques, such as SEM, often require fairly significant samples sizes (Kline, 2011) and the required focus on group-level results may mask important individual or subgroup effects (Barclay-Goddard, Epstein, et al., 2009).

In investigating response shift, researchers have proposed and implemented a number of measurement techniques. However no single technique has become standard for the discipline. As with all research, the specific research problem should direct the methodology. Since response shift assessment has not been incorporated into most health care clinical trials, to best support the response shift integration into practice a technique that can be performed as a secondary analysis of existing data would be preferred. SEM, a statistical approach, meets that criterion.

SEM

Based on the exploration of the structure of means, variances, and covariances of variables of interest, health care researchers have used a statistical modeling technique, SEM, to explore response shift. This family of related multivariate analysis procedures includes path analysis, confirmatory factor analysis, and structural regression (Kline, 2011). SEM enables researchers to study and test complex linear relationships between both observed and latent (unobserved) variables and allows variables to be indicated by multiple measures. Primarily a confirmatory technique, the SEM process involves a researcher hypothesizing the causal model, depicting it as a path diagram, and then testing the model using empirical data. The flexibility of specifications and requirement

for theoretical justification enable SEM to be used to effectively to address and model the complexity of health care and response shift.

SEM Overview

SEM is a collection of multivariate analysis procedures that focus on means, variances, and covariances to explore relationships between observed and latent (unobserved) variables. Confirmatory factor analysis, a SEM model, is used to determine if collected data fit a theory-based measurement model. Based on covariance, the procedure is designed to accomplish two goals—to understand the patterns of covariance and to use the researcher-specified model to explain as much of the variation as possible (Kline, 2011). The variables that can be included in a structural equation model support the complexity of social science research. Variables may be observed or unobserved (latent) with some of the latent variables representing measurement error. The scale of a latent variable is arbitrary and must be set in the model by the researcher. By setting the variance of a latent variable to 1, the scale can be standardized. Alternatively, the variable may take on the scale of one of its indicator variables (Lei & Wu, 2007). These conventions allow for model simplification since when fixed in either manner, the variables are not estimated from the data. Similarly, since for endogenous variables all effects are included in the model, no unanalyzed associations occur between these variables (Kline, 2011).

Kline (2011) identified six steps typical of SEM—specify the model, confirm model identification, select measures and collect data, estimate the model, respecify the model, and report the results. In the first step of the process, researchers create a

hypothesis of the relationships of interest and their interconnections based on literature, observations, and logical reasoning. A good model should be theoretically justifiable, simple and straightforward, and reproduce the correlation matrix based on the constraints. In the identification step, the researcher verifies that it is theoretically possible for a computer to derive a unique estimate for every model parameter. The third step of this technique is to collect the data and prepare the dataset. Preparation and screening of the data will ensure that if no SEM solution is produced, the null result is a function of an invalid model and not data issues. With the data available and verified, the researcher uses a SEM computer tool to conduct the analysis. This step, called estimation, includes evaluating model fit, interpreting the parameter estimates, and considering equivalent (and near-equivalent) models. The chi-square test can be used to test the null hypothesis; however, this test is sensitive to sample size and may reject a reasonable data fit based on a large number of samples (Lei & Wu, 2007). Alternative goodness-of-fit indices have been created to adjust for this effect and SEM experts recommend that multiple indices be considered when overall model fit is being evaluated (Kline, 2011; Lei & Wu, 2007). The fifth step in the SEM process is respecification, the reworking of the model to address any issues identified in earlier steps (Kline, 2011). A key consideration of respecification is that changes should be guided by rational considerations and Kline recommended that researchers identify theoretically justifiable changes in the original Step 1 model specification. The final step in the SEM research is to document the analysis completely and accurately. Building on the earlier steps of the SEM process, reported results should include a comprehensive review of the specification of the model,

documented validation of identification, complete characterization of the sample data, listing of SEM program and assumptions, review of the estimation, and discussion of any respecification required. Confirmation bias should be addressed and the implications of the analysis, whether the model was retained or not, should be highlighted.

Due to its focus on covariance and since standard error may not be accurate with small samples, SEM is a large sample technique. This requirement can have significant impact on the research questions that can be addressed. Because SEM can produce both simpler and more complex models, there is no universal guideline for sample size. Researchers have proposed that the ratio of cases to the number of model parameters be set at between 10 and 20:1 (Kline, 2011). However in practice the typical sample size for SEM studies is about 200 cases, the approximate median sample size of peer-reviewed articles published in psychology and management science journals that used SEM techniques. The flexibility available from SEM can provide researchers with an excellent tool for understanding complex social situations including response shift. A number of QoL researchers including Hamidou et al. (2011), Oort (2005b), and Schwartz et al. (2011) identified SEM as the most pragmatic method of integrating response shift into clinical practice since it addresses multiple variables simultaneously and it does not increase effort or time commitment from the subject or clinical investigator during data collection.

SEM for Response Shift

While a variety of factor analysis and covariance methods have been proposed, the use of SEM has emerged as a promising response shift assessment technique. Two

methods have been demonstrated in the literature, the Schmitt method and the Oort method. While both SEM, there are differences in both the theory and methodology. The Schmitt method begins with an unconstrained model and adds constraints, retaining only those constraints that work. Conversely, the Oort method begins with a fully constrained model and releases constraints that are untenable. Ahmed et al. (2009) directly compared the Schmitt and Oort SEM techniques in a COPD population that participated in a self-management program. Using two QoL instruments and data from baseline and 1 year, the Schmitt method did not identify any response shift while the Oort method found significant changes. The authors concluded that these subjects did experience response shift that underestimated change in physical health and that the Oort procedure was more sensitive in detecting response shift than the Schmitt method.

The body of work to support the Oort SEM methodology for response shift assessment includes publications that address the theory, application of the SEM technique to clinical data, and mathematical correlations to recalibration, reprioritization, and reconceptualization (Oort, 2005a; Oort, 2005b; Oort et al., 2005). Ahmed et al. (2005) investigated response shift in a post-stroke population at 6 months, comparing three techniques—SEM, the then-test, and an individualized approach. The results were mixed with the then-test and individualized approach identifying response shift and the SEM not showing any response shift. Visser et al. (2005) also compared three methods of assessing response shift in a cancer study with QoL collected at baseline and 3 months post-surgery. The methods included the then-test, a preference-based approach, and SEM. In the study, all three methods identified response shift with the then-test and SEM

results largely comparable. The preference-based approach also identified response shift but in different domains and in a divergent direction. The authors concluded that due to the limitations of the preference-based test, the time required and the qualitative nature of the data, the then-test and SEM approaches were preferred for future research.

King-Kallimanis et al. (2009) explored the difference in measurement and conceptual perspectives of response shift by applying SEM to a cancer population undergoing surgery. Five measurement response shift biases were identified but no explanation response shifts were identified. This application successfully added the exogenous factors of cancer site, health status, sex, age, optimism, and social comparison to the SEM methodology. In a further expansion of the Oort method, Barclay-Goddard, Lix, et al. (2009) evaluated response shift over multiple occasions in a post-stroke population. Response shift was identified at both 6 and 12 months though the shifts were not identical at these timepoints. The authors validated the SEM modeling technique for response shift and illustrated the usefulness of the information. Schwartz et al. (2011) compared three statistical methods for evaluating response shift, including SEM. Only small response shift effect sizes were identified in the multiple sclerosis population by all methods. However, the research did support the operationalization, interpretability, and data usage of the techniques. While based on the range of results a definitive best method could not be selected, in the comparison SEM yielded the clearest findings, effectively compared disease-specific to generic outcomes, and was determined to be the most successful.

When applied to response shift and health care research, SEM strengths include the ability to incorporate multiple variables into the model, the inclusion of measurement error, and the fact that this technique can be applied without additional data collection. Finally, since in the social sciences the magnitude of the effect is often most important and not the specific result of the statistical test, SEM provides better estimates of effect size for observed variables than many other mathematical techniques (Kline, 2011). A detailed methodology for this research study based on the Oort SEM technique is included in Chapter 3.

Conclusion

The use of patient-reported outcomes to support clinical trials and evidence-based medicine have highlighted the potential impact response shift phenomenon can have on study interpretation and conclusions. When individuals report their QoL at different timepoints, the results may be affected by changes in the individual's internal standards, values, or conceptions. This can complicate the evaluation of longitudinal data and result in under- or overestimation of treatment effects. In this review I have outlined the theoretical foundation of response shift and presented current research in health care including the diversity of findings in specific medical conditions and over different timeframes. Researchers have explored a variety of techniques to identify and quantify response shift including retrospective rating, individualized approaches, and statistical methods. While the scholarly literature has explored response shift, the significance in comparative clinical trials of medical devices has not been extensively studied.

Investigating the clinical significance in a randomized clinical trial comparing two spine interventions, a population expected to experience response shift, would support the translation of response shift research into practical guidance for investigators and clinical trial designers. By using SEM, a statistical method sensitive to all components of response shift and appropriate for secondary analysis, any differences in response shift impact on the control and test groups at different timepoints could be explored. A practical application of response shift assessment would support clinical investigators, health care professionals, and QoL researchers in determining if response shift analysis should be incorporated into comparative trials using patient-reported outcomes. Chapter 3 contains the specifics of the study design, clinical data, and modeling techniques to investigate response shift in the ISISS spine intervention study (VertiFlex, 2013).

Chapter 3: Research Method

In spine interventions, clinical investigators typically compare subject-reported QoL scores from pre- and postintervention to document the effectiveness of treatments. However, subjects can adjust the way they score their QoL based on factors unrelated to the intervention including changing individual internal standards, values, and prioritizations. These changes, called response shift, could invalidate direct comparison of the QoL scores. Using data from a spine intervention trial and applying SEM techniques, I used a quantitative study design to explore response shift phenomenon and analyze the potential impact on comparative clinical trial data interpretation. Evaluating changes in response shift over time and determining if there was a difference in response shift between treatment groups will provide researchers with insight into the clinical significance of response shift for this subject population. This chapter contains an outline of the research design, data preparation, study population, the data analysis plan, and modeling framework for this research.

Research Design

Using data from a randomized clinical trial comparing the effectiveness of two medical devices and SEM, I explored response shift through 12 months using a longitudinal confirmatory modeling research design. Data from four lumbar spinal stenosis and spine QoL and function questionnaires were used to address the following research questions:

1. Do treated back pain patients experience a difference in response shift between baseline and 3 months and between baseline and 12 months postintervention?
2. Does response shift phenomenon influence the clinical comparison of patient-reported outcomes between baseline and 12 months in a randomized clinical trial for a spine intervention?

Answering these research questions supported understanding of response shift in a variety of ways. First, by exploring response shift over time, I provided information about the existence of response shift and potential catalyst events in a spinal intervention population. If response shift was not detected, then spine investigators could continue to implement standard research designs with an increased level of confidence that their results were not biased by this phenomenon. If response shift could be detected, then investigators would be aware of this potential bias and could incorporate this knowledge into their clinical trial data interpretation. I supported the identification of catalyst events by characterizing differences in response shift between postintervention and later occasions. By addressing the second research question, I focused on clinical trial comparisons and response shift. To minimize bias, researchers often prefer study designs in which subjects are blinded to their treatment. However, in medical device clinical trials involving surgery, it is often difficult to prevent subjects from knowing which intervention they received. This knowledge has an unknown impact on expectations and therefore could influence response shift. Determining if response shift impacted two treatment groups differently could influence the confidence investigators have in their

study conclusions. Finally, performing this research using SEM supports the assessment of the practicality of implementing this methodology for response shift investigation.

The ISISS study, the source of the data, was designed to determine the equivalence between the control and investigational devices (Treatments A and B) by comparing the number of successful interventions at 24 months. Researchers defined a successful intervention based on the improvement in patient-reported outcomes collected by the ZCQ, ODI, SF-12, and VAS for back, right leg, and left leg pain. Investigators focused secondary endpoint analysis on comparisons of patient health status pre- and postintervention using clinical difference thresholds and *t* tests. However, these statistical comparative techniques assume that at every timepoint subjects used the same internal standards, values, and conceptualization to assess their QoL. QoL researchers have identified that response shift phenomenon may invalidate this assumption but that QoL instruments are not designed to determine the impact of response shift (Razmjou et al., 2009; Schwartz et al., 2006; Sprangers & Schwartz, 1999; Wilson, 1999). Therefore, an alternative analysis technique was required. Through my analysis of the literature, I identified SEM, an advanced modeling technique, as an effective tool for investigating the phenomenon and employed it for this research.

Modeling techniques have been used effectively to create theory, to describe cause-and-effect relationships between variables, and to predict system output based on inputs. In an alternative application of modeling, Oort et al. (2005) demonstrated that evaluating variable invariance in a SEM model could detect and provide insights into response shift. After an appropriate measurement model was developed, constraints were

added and removed systematically. The fit and equivalence of the resulting models were evaluated. If the constraints significantly changed the model's goodness of fit, then response shift associated with the constrained variables was identified and could be investigated in more detail.

In this research, I created a theoretically justified structural equation model using subject-reported QoL responses to validate and optimize the model. As SEM terminology replaces dependent and independent variables used in other statistical methods with observed and latent variables, I incorporated these variables in a confirmatory path model to represent subject QoL reporting. I selected the observed variables from the questionnaire data provided from the QoL instruments and designated the latent variables, the unmeasurable true physical and mental states of the respondents, as PQoL and MQoL. Due to inherent measurement error and the many ways QoL concepts are interpreted by individuals, latent variables cannot be directly collected by behavior instruments. When an adequate model was supported, I systematically adjusted the constraints and analyzed the results using LISREL 9.1 SEM software (Scientific Software International, 2013). To determine if response shift was present, I evaluated the invariance of the model, direct effects, and variable responses. In order to address the research questions, I tested the constraints associated with measurement occasions (3 months, 12 months) and the exogenous variable treatment group (A or B). Further details of the analysis are outlined in the Data Analysis Plan section.

I selected a statistical approach, SEM, for this research to maximize the generalizability and value to health care clinical investigators. Other alternatives, such as

retrospective ratings and individualized approaches, require response shift investigation to be included in the original study design as an additional input. Because of the added research complexity and subject burden, few current comparative clinical trials have incorporated this data collection into their research. Statistical techniques only require data from QoL instruments used in the primary analysis and so can be effectively used for secondary analysis. Additionally, SEM addresses both measurement and conceptual issues in a single model, can identify all three components of response shift, and can address multiple follow-up timepoints (Hamidou et al., 2011; Oort et al., 2009; Schwartz et al., 2011). This method has also been demonstrated in the literature (Barclay-Goddard, Lix et al., 2009; King-Kallimanis et al., 2011, King-Kallimanis et al., 2012; Oort et al, 2005). For these reasons, I identified a SEM technique using data from an on-going spine intervention study to support this investigation.

Data and Instrumentation

Statistics & Data Corporation (SDC), the ISISS study sponsor's data manager, at the direction of the sponsor, VertiFlex, Inc., provided me with the limited dataset in an Excel format. VertiFlex obtained permissions to use all instruments as part of the ISISS study including authorization for secondary analysis (Appendix A). As the ISISS study was designed and conducted to support a premarket application for the FDA, the sponsor employed quality control techniques to ensure the validity of the data including (a) electronic case report forms (eCRFs) that included built-in edit checks, (b) on-site monitoring, and (c) general quality control. Based upon the quality control measures in place, I accepted the assumption that data accuracy was sufficient to support this

research. The research dataset was formatted as an Excel file. I will maintain the limited dataset provided by VertiFlex for audit purposes in compliance with Walden University policies.

Instruments

Four primary lumbar spinal stenosis and spine QoL and function questionnaires were used to collect data for the ISISS study—the ZCQ, ODI, SF-12, and VAS for back, right leg, and left leg pain. Each of these instruments was previously validated in spine and back conditions. The details of the validations can be found in the literature and are outlined in the individual instrument descriptions that follow.

ZCQ. The ZCQ is a condition-specific instrument for lumbar spinal stenosis. It is a self-administered three-section patient survey covering symptom severity, physical functioning, and patient satisfaction with treatment. Satisfaction with treatment, Part 3, is only scored after an intervention has been performed. Items are scored with a 5-point Likert scale for symptom severity and a 4-point Likert scale for physical function and satisfaction. The results are calculated by subscale and expressed as a percentage of the maximum possible score with higher scores representing increased disability or dissatisfaction. As the result for each subscale is expressed as a percentage of the maximum possible based on questions answered, missing answers have been addressed in the scoring. Validity, reliability, and predictive ability have been well studied (Pratt, Fairbank, & Virr, 2002; Stucki, Daltroy, Liang, Fossel, & Katz, 1995; Stucki et al., 1996)

ODI. The ODI is a condition-specific instrument for spine disorders and low back pain. Composed of 10 items, the self-administered questionnaire is designed to

evaluate how spine issues, specifically back or leg pain, are impacting the respondent's ability to handle everyday life. Items are scored on a 6-point Likert scale with the final score calculated as a percentage with lower scores representing minimal disability and higher scores increased disability. As one of the most commonly used outcome measures in spine issues, validity, reliability, and predictive ability have been well studied (Fairbank, Couper, Davies & O'Brian, 1980; Fairbank & Prysent, 2000; Pratt et al., 2002; Walsh et al., 2003).

SF-12. The SF-12 is a multi-purpose health survey that measures functional health and well-being using 12 questions. Commonly called the SF-12, this instrument does not target a specific population or disease and measures multiple health domains including physical and mental components. Respondents report on physical and social activities that can be accomplished, how often they are performed, and the level of difficulty associated with them. The SF-12 is an adaption of the longer 36 question SF-36 survey and was designed to be easier and faster for patients to complete. This instrument is recommended for large studies and for group comparisons; however the more granular SF-36 is preferred for individual decision-making (Resnik & Dobrzykowski, 2003)

The SF-12 scale includes two primary subscales, a physical composite subscale (PCS) and a mental composite subscale (MCS), and eight domains. The subscales and domains are generated by scoring, combining, weighting and normalizing test item responses. The specific scoring formula is proprietary, however the developer, Quality Metric, provided the subscale and domain results as part of the licensing agreement. The scales have a range of 0 – 100 with norms set by the developer based on a mean of 50 and

standard deviation of 10. Norms therefore vary and are uniquely associated with population age groups. In the U.S. general population, PCS normal decreases with age and MCS normal increases with age. The physical score is made up of four individual domains—Physical Function (PF), Role Physical (RP), Bodily Pain (BP), and General Health (GH). The mental score is also composed of four domains—Mental Health (MH), Role Emotional (RE), Social Functioning (SF), and Vitality (VT). The validity and reliability of the SF-12 subscales and domains have been well established for both general and low back pain patient populations (Jenkinson et al., 1997; Luo et al., 2003; Resnik & Dobrzykowski, 2003; Ware, Kosinski & Keller, 1996).

VAS. The VAS is a single item measurement tool for patient pain severity. Respondents report their pain by marking their pain level on a 100 mm line. No pain is the anchor at the left extreme and worst pain possible is the scale of the right extreme. The score is determined by measuring the number of millimeters between the left starting point and the patient's mark. Outcomes are between 0 and 100 with higher scores indicating increased pain. The VAS is the most frequently used pain outcome measure for back pain and validity and reliability have been demonstrated (Litcher-Kelly, Martino, Broderick, & Stone, 2007; Moore, Moore, McQuay, & Gavaghan, 1997; Olaogun, Adedoyin, Ikem & Anifaloba, 2004).

Data Collection and Preparation

I requested a limited dataset containing demographic and QoL information from the ISISS study at an in-person meeting with Steve Reitzler, Vice-President Regulatory and Clinical Affairs, VertiFlex, Inc. After input from Western IRB who determined that

the dataset met all the criteria for a deidentified dataset and that no additional approval from the subjects was required for secondary analysis, Mr. Reitzler directed the ISISS data management provider, Statistics & Data Corporation (SDC), to e-mail a dataset to me for use in my dissertation (Appendix A). To further deidentify the data and ensure no negative impact on the ISISS study, subject identification codes were converted from a site-subject format to a subject only format, treatment group was blinded, and only 12-month and earlier follow-ups were included in the data.

As of May 2012, 476 subjects had been enrolled in the ISISS study with 288 subjects having received an implant and reached 12 months post device surgery. To support the structural equation model, only subjects who had data from the 12-month follow-up were included in the analysis dataset. The data collected in the ISISS study, provided in the limited dataset, and required for the research dataset are outlined in Table 2.

I created a raw data file limited to all subjects who met the inclusion criteria. I ensured the variables could be implemented by the LISREL 9.1 software and compared the source and LISREL datasets to each other to ensure accurate transfer and formatting. After error checking was complete, I saved the file and made backups. Backups were stored separately and as part of an offsite backup service (Carbonite, Boston, MA). The dataset was screened for extreme collinearity, outliers, missing data, and multivariate normality in accordance with Kline (2011) and LISREL 9.1 program guidelines.

Table 2

Data Collection for the ISISS Study and Research Data Set

Data Collected	ISISS Study Scheduled Data Collection	Limited Data Set	Research Data Set
Demographic Info	Baseline	Baseline	Baseline
Randomization	Surgery	Surgery	Surgery
Spine X-rays	Baseline, Discharge, 6 weeks, 3, 6, 12, 18 & 24 months	-	-
Neurological status	Baseline, Discharge, 6 weeks, 3, 6, 12, 18 & 24 months	-	-
Adverse events (hospitalizations, death)	Baseline, Discharge, 6 weeks, 3, 6, 12, 18 & 24 months	-	-
QoL Instrument – Patient Satisfaction	6 weeks, 3, 6, 12, 18 & 24 months	-	-
QoL Instrument - ZCQ	Baseline, 6 weeks, 3, 6, 12, 18 & 24 months	Baseline, 6 weeks, 3, 6 & 12 months	Baseline, 3 & 12 months
QoL Instrument - ODI	Baseline, 6 weeks, 3, 6, 12, 18 & 24 months	Baseline, 6 weeks, 3, 6 & 12 months	Baseline, 3 & 12 months
QoL Instrument - SF-12	Baseline, Discharge, 6 weeks, 3, 6, 12, 18 & 24 months	Baseline, Discharge, 6 weeks, 3, 6 & 12 months	Baseline, 3 & 12 months
QoL Instrument - VAS	Baseline, Discharge, 6 weeks, 3, 6, 12, 18 & 24 months	Baseline, Discharge, 6 weeks, 3, 6 & 12 months	Baseline, 3 & 12 months

Modeling Dataset

I included the variables listed in Table 3 in the research dataset for use in assessing and optimizing the SEM model. In accordance with Kline's (2011) recommendation that scales be used as source data for SEM and as these outputs would be more representative of the results used for clinical interpretation, I did not include individual item responses in the dataset.

Table 3

Modeling Dataset Variables

Variables	Description
Age	in years at time of surgery
Race	
Gender	
Vertebral levels	1 level or 2 consecutive levels
BMI	
Treatment group	coded as A or B - investigational device (VertiFlex Superion™) or control device (X-Stop®). The assignment of code to specific treatment has been blinded.
Occasion	Baseline, 3-month follow-up, 12-month follow-up
ZCQ part 1 score	Symptom severity
ZCQ part 2 score	Physical function
ZCQ part 3 score	Satisfaction with treatment - for follow-up intervals only
ODI score	
PCS	Physical Component Subscore of SF-12
MCS	Mental Component Subscore of SF-12
PF	Physical function domain of SF-12
RP	Role physical domain of SF-12
BP	Bodily pain domain of SF-12
GH	General health domain of SF-12
V	Vitality domain of SF-12
SF	Social functioning domain of SF-12
RE	Role Emotional domain of SF-12
MH	Mental health domain of SF-12
VAS back pain score	
VAS leg pain score	Right leg and left leg scores added together to create one variable

Note. BMI = Body Mass Index; ZCQ = Zurich Claudication Questionnaire; ODI = Oswestry Disability Index; SF-12 = Short form general health survey; VAS = Visual Analog Scale for pain.

Study Population

This study involved secondary analysis of data collected to support a pre-marketing approval of a minimally invasive spinal implant, the ISISS study (VertiFlex, 2013). Data were collected from June 2008 through May 2012.

ISISS Population

The ISISS study used a prospective, randomized, multicenter trial design to compare the safety and efficacy of a minimally invasive spine implant for treatment of lumbar spinal stenosis (Loguidice et al., 2011; VertiFlex, 2013). As a degenerative disease, lumbar spinal stenosis symptoms can only be treated but not cured. Therefore, the primary goal of treatment is to maximize function and maintain QoL. The ISISS study recruited subjects from 32 sites in the United States. Clinical investigators identified potential subjects from their patient populations who met study inclusion criteria of being at least 45 years old, had experienced a minimum of 6 months of moderate spinal stenosis symptoms, and been unresponsive to other treatments. These patients were presented with the opportunity to enroll in the ISISS clinical trial (VertiFlex, 2013). If they agreed and after signing informed consent, subjects were randomized to either the control or the investigational device groups with half the subjects assigned to receive the control device (X-Stop®) and half to receive the investigational device (Superion™). The subjects had the surgery performed and the device implanted. Subjects returned to the investigator's sites for follow-up and completion of QoL questionnaires at regular intervals through 24 months. The data from

these follow-ups were collected in 21 CFR part 11 compliant data systems that included electronic case report forms with built-in error checks.

Sample Size

The ISISS study used a non-inferiority hypothesis between the two treatment groups and a Bayesian adaptive approach for sample size selection with the final number of subjects enrolled equal to 463. As of May 2012, when the dataset was created, 288 subjects had completed their 12-month follow-up visit. I combined the data into a single data file suitable for import into LISREL 9.1, a SEM software package.

Kline (2011) instructed that sample size for SEM is dependent on the specifics of the model so no universal guidelines exist. The number of parameters in the final model, specific estimation algorithm, and distributional characteristics of the data all impact required sample size. Researchers have proposed that the ratio of cases to the number of model parameters be set at between 10:1 and 20:1. In this research, the model used to assess response shift had a ratio of 18:1 cases to parameters. However, this guidance is a rule of thumb and cannot be rigorously tested since the details of each model and data characteristics have a significant impact on power estimates. In practice the typical sample size for SEM studies published in peer-reviewed literature is about 200 subjects (Kline, 2011). Using a representative model with parameter numbers similar to the model used in this research, Oort (2005b) calculated the sample size required to detect reprioritization change. Reprioritization represented the sample size required for the Oort SEM analysis as recalibration and reconceptualization have a larger impact on observed means and covariances and therefore require fewer cases. Oort reported that for a

statistical power of 80%, a sample size of 170 was required and to increase the power to 90%, a sample size of 228 was sufficient. Therefore, a sample size of 263 complies with SEM established practice and was adequate to identify SEM tested response shift for my research.

Data Analysis Plan

Using SEM techniques and LISREL 9.1 modeling software, I investigated response shift in QoL data reported in the ISISS study to address the following research questions:

1. Do treated back pain patients experience a difference in response shift between baseline and 3 months and between baseline and 12 months postintervention?
2. Does response shift phenomenon influence the clinical comparison of patient-reported outcomes between baseline and 12 months in a randomized clinical trial for a spine intervention?

This analysis plan contains an overview of the SEM methodology, study hypotheses with acceptance criteria, an explanation of the initial model, and model respecification alternatives.

SEM Process

To investigate response shift, I implemented the three-step SEM process described by King-Kallimanis et al. (2010):

1. Establish a measurement model.
2. Test invariance across measurement occasions.

3. Add exogenous variables and test direct effects.

Step 1: Measurement model. In the first step I developed a confirmatory factor analysis (CFA) model that was an appropriate measurement model with good fit and clear interpretation of the data. As recommended by Kline (2011), goodness of fit was evaluated using two indicators—chi-square and root mean square error of approximation (RMSEA). The chi-square value represents the equivalence of the model predicted means, variances, and covariances compared to the observed means, variances, and covariances and supports the assessment of a model's overall goodness-of-fit. In general, a good fit is indicated if the chi-square value is not significant. However since chi-square values are a measure of exact fit, the results can exhibit significant sample size sensitivity where in large samples even very small differences may be identified as significant. Adding RMSEA, an index of approximate fit, to the evaluation overcomes this limitation. RMSEA of < 0.08 was interpreted as a reasonable fit and < 0.05 as a close fit (Barclay-Goddard et al., 2009; King-Kallimanis et al., 2011).

When initial fit was not acceptable, the model was respecified according to theoretical associations outlined later in this chapter (Respecification Alternatives). I repeated the respecification process until I obtained a model with reasonable fit values.

Step 2: Invariance across measurement occasions. Next, I used the SEM model and a two-stage process to evaluate response shift. In the first stage, I assessed overall response shift. Then I evaluated the model to detect the specific response shift components of reconceptualization, reprioritization, and recalibration.

Step 2, Stage 1: Overall response shift assessment. In the first stage, I assessed invariance of the model across the study follow-up timepoints by constraining all factor loadings and intercepts to be equal across baseline, 3-months, and 12-months. Using a chi-square difference test, the new model was compared with the model from Step 1. The chi-square difference statistic tests the hypothesis that two nested or hierarchical models are equal (Kline, 2011). A significant result indicates the models are not equal and the model with more free parameters fits the data better. Hierarchical models are those where one model is related to the second model only by the addition or elimination of free parameters. A significant chi-square difference result between the unconstrained (step 1) and constrained (step 2) models indicates response shift (King-Kallimanis et al., 2010; King-Kallimanis et al., 2011; King-Kallimanis et al., 2012).

To expand on this conclusion, recall that latent variables represent the true value of the unmeasured attributes (i.e., wellbeing associated with physical aspects of life, PQoL). Based on data covariance and means matrices, the unconstrained model (final model from Step 1) provides a valid representation of the importance and contribution of the observed variables (pain, function) on the latent variable. Oort's (2005b) response shift methodology focuses on these relationships. When multiple occasions are included in the model, the use of validated instruments supports the assumption that the relationship between the observed variables and the latent variables remains the same across all occasions. For example, if pain explains 60% of PQoL at 3 months, pain should also explain 60% of PQoL at 12 months even if the actual observed values of pain are different at these two occasions. So when the variable relationships are the same at every

occasion, changes in latent variable means will be fully explained by changes in the observed variable means. Therefore, mathematically adding this assumption to the model by including across occasion equality constraints should not significantly decrease the fit of the model. When assessing the invariance of SEM models, Oort (2005b) identified response shift when models were not equivalent and no response shift when the models were equivalent. Therefore, a chi-square difference test that supports that the unconstrained and constrained models are not equivalent also supports the finding of response shift in the observed data.

Step 2, Stage 2: Evaluation of response shift components. The second stage of invariance testing requires the sequential removal of constraints on individual variables to assess the response shift components and detect their location (King-Kallimanis et al., 2011; King-Kallimanis et al., 2012; Oort, 2005b). Working in iterative series addressing each observed variable separately, I removed the equality constraints on factor loadings and intercepts at all measurement occasions and assessed the fit using the chi-square difference test. A significant chi-square difference was interpreted as response shift in the associated parameter (Barclay & Tate, 2014; Oort, 2005b). To guard against family-wise errors, I based significance of the chi-square difference test on Bonferroni adjusted levels of significance (King-Kallimanis et al., 2011; King-Kallimanis et al., 2012). For each series, the freed variable that was both significant and produced the largest chi-square difference was retained to create an improved model. The process was repeated until the model's goodness of fit could not be further optimized.

Oort (2005b) described and supported relationships between the three components of response shift and changes in structural equation models. These relationships are outlined in Table 4 and were used to guide the response shift component evaluation.

Table 4

Oort SEM Correlation to Response Shift Components

SEM findings over intervals	Response shift component
Factor-loading patterns	• Reconceptualization
Factor-loading magnitude	• Reprioritization
Residual factor variances	• Recalibration - nonuniform
Intercept changes	• Recalibration - uniform

In Step 2, Stage 2, I evaluated the model components based on the Stage 1 optimized model and assessed response shift in the following order (a) reconceptualization, (b) reprioritization, and (c) recalibration. In accordance with Oort (2005b), reconceptualization analysis was based on the evaluation of the pattern of factor loadings. Changes from zero to non-zero or positive to negative between measurement occasions identified reconceptualization response shift. Reprioritization was based on a comparison of the magnitude of factor loadings across occasions. Using standardized factor loadings, I calculated the difference between each set of measurement occasions by subtracting one from the other. Standardized factor loadings are correlation estimates between the observed and latent variables and when squared (R^2) are proportions of explained variance. Consistent with King-Kallimanis et al. (2011) and SEM standard practice (Kline, 2011), I interpreted a difference in factor loadings of 0.10 as significant. Similarly, to identify recalibration, changes in residual factor variances (nonuniform

recalibration) and intercepts (uniform recalibration) across measurement occasions were assessed.

The requirement to characterize response shift components was guided by the specific hypotheses tested in this research. Since the identification of at least one component of response shift resolved the hypothesis, I ended the identification of components after this occurred.

Step 3: Add exogenous variables. In the final step, I tested exogenous variable invariance and investigated direct effects on the observed indicator variables. First, I added the exogenous variable of interest, treatment group, to the model and assessed the model for goodness of fit. Treatment group was free to correlate with PQoL and MQoL while all direct effects were fixed to zero. To test observed variable invariance, I created a series of models where the relationship between treatment group and the observed variables were individually freed. A significant chi-square difference test and a parameter change of at least 0.10 suggested a lack of invariance and a potential direct effect of the exogenous variable on the observed variable (King-Kallimanis et al., 2010; King-Kallimanis et al., 2011; King-Kallimanis et al., 2012). In each series, the parameter associated with the largest significant improvement denoted a direct effect and was left free to be estimated. I repeated the process until no further significant direct effects were identified.

To identify potential chance findings and remain consistent with the methods of other response shift researchers, I repeated the assessment for direct effects on a model that included all available exogenous variables that could induce bias in QoL scores. The

exogenous variables added were treatment group, age, gender, number of vertebral levels treated, and BMI. I repeated the evaluation for direct effects using the new model. As the purpose of this testing was to gain additional insights, I used this information to further explain the results of the first Step 3 modeling assessment and not for hypothesis testing.

Hypotheses

$H_0$1: Response shift at 12 months is not different from response shift at 3 months.

$RS_{12} = RS_3$

H_a1: Response shift at 12 months is different from response shift at 3 months.

$RS_{12} \neq RS_3$

where

RS_{12} is response shift of study population between baseline and 12 months, and

RS_3 is response shift of study population between baseline and 3 months.

I evaluated the difference between response shift at 3 months and 12 months based on the methodology outlined in SEM process, Step 2. After a SEM model with good fit was determined from the research data (Step 1), the invariance of the model with respect to measurement occasions was tested (Step 2, Stage 1) and response shift components evaluated (Step 2, Stage 2).

For Stage 1, I added equality constraints and reassessed the model for goodness of fit using LISREL software. I evaluated the results based on the following criteria

- If the chi-square difference test between the constrained and unconstrained models was not significant, the two models were interpreted as equivalent. If

the models were equivalent then no response shift across any timepoints was identified and the null hypothesis was retained.

- If the chi-square difference test between the constrained and unconstrained models was significant, the models were not equivalent and response shift was identified. However, since this test detects response shift over the entire model (all three measurement occasions), additional analysis was required to assess if there was a difference between the two timepoints of interest (the 3-month and 12-month occasions). If required, I performed this Stage 2 analysis.

For Stage 2, I tested the model as described in SEM process, Step 2 and assessed the components of response shift based on Oort's (2005b) SEM correlation to response shift components (Table 4). The null hypothesis was rejected if I identified any component of response shift between the 3- and 12-month occasions. The null hypothesis was retained if I identified no response shift between the 3- and 12-month occasions. Response shift identified between other occasions was not used to address the hypothesis. I discontinued evaluation of response shift once the hypothesis was addressed.

The three components of response shift (reconceptualization, reprioritization, and recalibration) were assessed as follows. To determine reconceptualization, I reviewed the factor loading patterns across occasions in the final model and evaluated the results based on the following criteria.

- If the pattern of factor loadings between 3 months and 12 months changed (i.e. from non-zero to zero), reconceptualization was identified.

- If the pattern of factor loadings did not change, reconceptualization was not identified.

For reprioritization, I calculated the difference in factor loadings across occasions in the final model and evaluated the results based on the following criteria.

- If the factor loading difference between occasions was at least 0.10, reprioritization was identified in that parameter.
- If the factor loading difference was less than 0.10, reprioritization was not identified.

For recalibration, I calculated the difference in residual factor variances across occasions in the final model and evaluated the results based on the following criteria.

- If the residual factor variance difference between occasions was at least 0.10, nonuniform recalibration was identified in that parameter.
- If the residual factor variance difference was less than 0.10, nonuniform recalibration was not identified.

$H_0$2: Response shift found in the patient-reported outcome results for treatment group A at 12 months is not different from the response shift in treatment group B at 12 months.

$$RS_A = RS_B$$

H_a2: Response shift found in the patient-reported outcome results for treatment group A at 12 months is different from the response shift in treatment group B at 12 months.

$$RS_A \neq RS_B$$

where

RS_A is response shift of treatment group A at 12 months, and

RS_B is response shift of treatment group B at 12 months.

I evaluated the impact of response shift between Treatment A and Treatment B subgroups at 12 months based on the methodology outlined in SEM process, Step 3. With treatment group added as an exogenous variable, the model was assessed for goodness of fit and evaluated for direct effects. I used the following criteria to assess Hypothesis 2.

- If the optimized model included a direct effect for the exogenous variable treatment group that varied between measurement occasions, the null hypothesis was rejected.
- If no direct effect was demonstrated or the direct effect was equivalent between measurement occasions, the null hypothesis was retained.

Model Parameters

The first step of the SEM was creating a confirmatory factor analysis model that associated latent and observed variables. For this research, I identified the latent, observed, and exogenous variables as follows:

Latent variables. Two latent variables were used in the SEM model, PQoL and MQoL. These variables are consistent with general QoL research and the design of the SF-12 (Oort et al., 2005).

Observed variables. Observed variables were provided by the QoL and function instruments collected in the ISISS study—the ZCQ, ODI, SF-12, and VAS for back and

legs. Continuous variables such as scales and domains were preferred over item responses in accordance with SEM best practices (Kline, 2011).

Exogenous variables. The primary exogenous variable was treatment group to support the testing of hypothesis 2. Additional exogenous variables (age, gender, number of vertebral levels treated, and BMI) were also evaluated to better understand potential demographic and surgical confounding factors.

Initial Model

Based on instrument design, literature sources (King-Kallimanis et al., 2011), and theory, I created an initial model that contained six observed variables for PQoL (ZCQ Parts 1 and 2, ODI, PCS of the SF-12, back VAS and leg VAS) and three observed variables for MQoL (ZCQ Part 3 for 3-month and 12-month follow-ups, Mental Health and Role Emotional domains of the SF-12) included. The measures and their scales are summarized in Table 5.

Table 5

QoL Indices Used in Initial SEM Model

Instrument	Item	Scale
ZCQ - Lumbar spinal stenosis specific	• Part 1: Symptom severity • Part 2: Physical function • Part 3: Satisfaction	0 – 100; % of maximum, higher scores indicate increased symptoms, decreased function, decreased satisfaction
ODI - Spine disorders specific	• Disability composite	0 – 100; % of maximum, higher scores indicate increased disability
SF-12 -Generic health profile	• PCS: Physical composite subscore	0 – 100; higher scores indicate better physical health
	• MH: Mental health domain • RE: Role emotional domain	0 – 100; higher scores indicate better mental status
Pain VAS - Pain severity measure	• Back pain severity	0 – 100; higher scores indicate increased pain
	• Leg pain severity – addition of right and left leg VAS scores to create one variable	0 – 200; higher scores indicate increased pain

The relationships between the latent variables and observed variables were based on the instrument design with physical and pain measures associated with PQoL and mental components associated with MQoL. The resulting path diagram is presented in Figure 2.

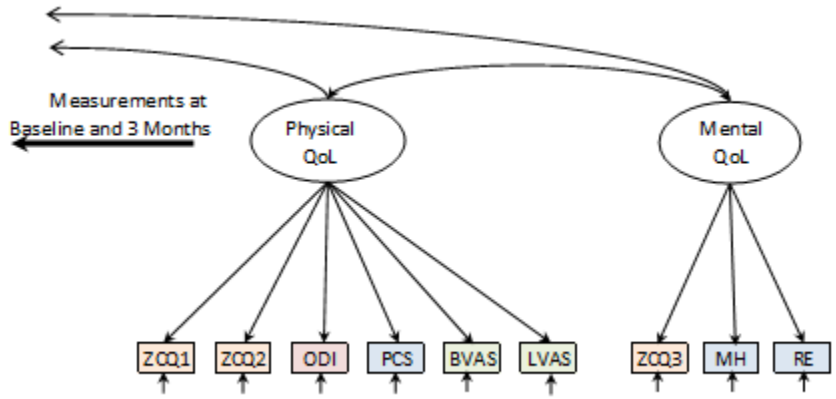


Figure 2. Path diagram of 12-month follow-up. ZCQ1 = Zurich claudication questionnaire-symptom severity, ZCQ2 = Zurich claudication questionnaire-physical function, ODI = Oswestry disability index, PCS = physical component of SF-12, BVAS = back visual analog scale, LVAS = leg visual analog scale, ZCQ3 = Zurich claudication questionnaire-satisfaction, MH = mental health domain of SF-12, and RE = role emotional domain of SF-12.

Since multiple intervals were included in the model, the final path model included three timepoints. The resulting initial model is represented by Figure 3.

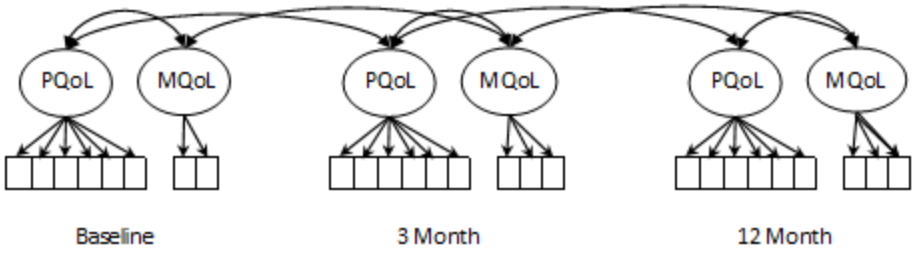


Figure 3. Path diagram for initial model. Note: Baseline timepoint has one less observed variable for Mental QoL as ZCQ3, Satisfaction with treatment, is not completed at baseline.

Respecification Alternatives

Model building was guided by theory and empirical results from previous research. In accordance with best practices for SEM, the following respecification

alternatives were documented prior to estimating the initial structural equation model. First, due to the potential for overlap in variables that measure either pain or physical function parameters, theory supports correlation between variables that would make the inclusion of multiple variables redundant. Kline (2011) recommended that redundant variables be removed from the model to support parsimony. If modification indices and standardized residuals support fewer indicators, the removal or substitution of indicators within the attribute groups in Table 6 would be supported by theory.

Incidental correlations when two indicators have more in common than the common factors are also theoretically justified (Oort, 2005c). Additionally, relationships may exist between theoretically related factors. King-Kallimanis et al. (2011) identified covariance in an SF-12 SEM model between mental health (MH) and role emotional (RE).

Table 6

Model Indicators Grouped by Attribute

Attribute	Indicators
Pain	ZCQ1 (symptom severity scored by ZCQ) BVAS (back pain scored by VAS) LVAS (leg pain scored by VAS) BP (bodily pain scored by SF-12)
Physical Function	ZCQ2 (physical function scored by ZCQ) ODI (disability scored by ODI) PCS (physical component score of SF-12) PF (physical function scored by SF-12) RP (role physical scored by SF-12) GH (general health scored by SF-12)

Another alternative to simplify the model would be to remove the indicators associated with pain. Symptom severity and pain scores, especially 12 months after an intervention and in an older population, may be attributed to factors other than spine issues. Some may be chronic while others transient, e.g. subjects overexerted by playing 36 holes of golf in the days prior to the follow-up visit. Therefore, pain may not be as relevant as physical function when evaluating back issues. Removal of the pain indicators would be theoretically justified if a simpler model was required.

Human Subject Protection and Research Validity

Walden University Institutional Review Board approved this research under IRB approval #05-08-14-0201068. As secondary analysis of data collected in a separate clinical trial, this research included no direct access to the human subjects. The original data continues to be maintained by VertiFlex and does not contain names or other subject unique identifiers, instead using unique subject screening numbers. For this research, the original screening number was replaced with a new identification number known only to VertiFlex to further enhance privacy protection. Human protection of subjects enrolled in the primary ISISS study was conducted in compliance with 21 CFR 56 where clinical investigators and the study sponsor, VertiFlex, verified that prior to study enrollment IRB approval was granted and all subjects signed informed consents. For this secondary analysis, I used the limited dataset only as permitted by this plan and as required by law and implemented safeguards to prevent inappropriate use or disclosure of the dataset. When the dataset is no longer required, I will destroy it in compliance with the Walden

University dissertation support requirements. I will also ensure that all Walden IRB requirements and conditions are met.

Threats to external validity, inappropriate generalization to populations not included in the research, have been mitigated by highlighting the research patient population as subjects in the United States who suffered from moderate lumbar spinal stenosis, who agreed to implantation of a minimally invasive medical device and were followed for 12 months. As health care literature has identified divergent response shift based on patient condition, outcomes, and follow-up interval, the results of this research may not be generalizable to other patient populations or other intervals. As secondary research based on experimental data from a randomized clinical trial, the design of the primary ISISS study addressed internal validity threats. Selection bias, differences in the study population, and experimenter bias have been mitigated by the use of required inclusion criteria and randomization. Additionally, instrument and study procedures were prescribed in detail and maintained throughout the study. Finally, the inability of the measurement instrument to measure the variables of interest, construct validity, has been mitigated by the use of advanced modeling techniques. SEM integrates multiple variables into the same model. Comparing changes in response shift and not characterizing the specific response shift, enabled any construct issues to be excluded from the analysis. Overall, while internal and external validity are always important, the design of this research has mitigated the potential impact of these research issues.

Summary

Changes in the internal standards, values, and priorities that patients use to assess health related QoL over time, response shift, could impact the results and conclusions of medical device clinical trials that use patient-reported outcome measures as primary endpoints. I used SEM to evaluate response shift using data from a randomized clinical trial of a spine intervention comparative study. The three-step SEM framework presented by King-Kallimanis et al. (2010) was used to evaluate response shift in the data. Overall, this chapter contained the research study design, study population, instrumentation, and a practical and effective SEM methodology to evaluate response shift in longitudinal comparative clinical trial data.

Chapter 4: Results

The purpose of this research was to explore the impact of response shift on clinical trial data interpretation at 1 year in an interventional spine clinical trial. Specifically, I investigated the difference in response shift experienced by patients between two measurement occasions. I also researched if the intervention the subject was randomly assigned impacted response shift in a way that could influence clinical interpretation. To investigate response shift, I developed a confirmatory path model and used it to test invariance over measurement occasions and the direct effect of patient characteristics. This chapter contains the results of the research including a data summary, the measurement model with goodness of fit results, and hypothesis testing.

Results

Population

From the ISISS study data collected between June 2008 and May 2012, I identified 288 subjects who had received a study interspinous medical device and had reached 12-month enrollment. Of these, 263 subjects reported 12-month QoL results. As the ISSIS study was designed to collect a representative sample of lumbar spinal stenosis subjects, this research subset should also be representative of this spine population. Data were limited only by 12-month follow-up data availability as I performed no sampling. The number of subjects unavailable (8.6%) was also consistent with the study expected loss to follow-up rate of 10%. The research population had an average age of 67.1 years, was 63% male, and 54% had two vertebral levels treated (Table 7). Ninety-seven percent of the subjects had data for all three timepoints (baseline, 3 months, and 12 months).

Table 7

Demographics

Variable	
Subjects (*n* =)	263
Age: Mean (SD)	67.1 (9.6) years
Race:	95% White
	2% African American
	2% Other
Gender	63% Male
	37% Female
Vertebral levels treated	46% 1 Level
	54% 2 Levels
BMI: Mean (SD)	29.7 (4.7)
Treatment Group	45% Group A
	55% Group B
Intervals	97% Baseline, 3- & 12-month
	3% Baseline & 12-month
Baseline ZCQ: Mean (SD)	3.28 (0.61) – ZCQ1
	2.67 (0.43) – ZCQ2
Baseline ODI: Mean (SD)	39.51 (12.82)
Baseline VAS: Mean (SD)	56.75 (25.84) – Back (max = 100)
	99.50 (48.39) – Legs (max = 200)
Baseline SF-12: Mean (SD)	28.53 (8.24) – PCS
	49.86 (13.05)– MCS

Note. ZCQ = Zurich Claudication Questionnaire; VAS = Visual Analog Scale for pain; ODI = Oswestry Disability Index; SF-12 = Short form general health survey; PCS = Physical Composite Subscale; MCS = Mental Composite Subscale.

I prepared the data using LISREL 9.1 software (Scientific Software International, 2013). To consolidate the Excel datasets into a single dataset, I separated the variables based on follow-up interval, converted ordinal and alpha data into numbers, and renamed labels to comply with LISREL software requirements. I imported the raw data into LISREL and ran the LISREL Statistics Data screening function to summarize the data. The LISREL standard full information maximum likelihood (FIML) estimation method was used to address missing data.

Across Occasion Response Shift

Hypothesis 1 testing required two steps—identifying an appropriate measurement model and testing for across occasion invariance as outlined in Chapter 3, SEM Process, Steps 1 and 2. To perform this analysis, I created and evaluated a series of models. Table 8 contains the association of observed variables with latent variables for these models and Table 9 presents the results of goodness of fit assessments.

Table 8

Observed and Latent Variables for Each Research Model

Model		PQoL	MQoL	Exogenous Variables	
1.0	Initial Model	ZCQ1, ZCQ2, ODI, PCS, BVAS, LVAS	ZCQ3, MH, RE	none	Figure 3
1.1	Initial Model with ZCQ3 removed	ZCQ1, ZCQ2, ODI, PCS, BVAS, LVAS	MH, RE	none	-
1.2F	Modified Model - Final	ZCQ2, BVAS, PF, BP, RE	MH, RE	none	-
2.1	Across occasion factor loadings - constrained equal	ZCQ2, BVAS, PF, BP, RE	MH, RE	none	-
2.2F	Factor loadings freed	ZCQ2, BVAS, PF, BP, RE	MH, RE	none	-
3.0	Model 2.2F plus Treatment group	ZCQ2, BVAS, PF, BP, RE	MH, RE	Treatment Group	-
3.1	Model 2.2F plus all exogenous variables	ZCQ2, BVAS, PF, BP, RE	MH, RE	Treatment Group, Age, Gender, Levels Treated, BMI	Figure 4

Note. F denotes final model for steps 1 and 2. PQoL = Physical Quality of Life; MQoL = Mental Quality of Life; ZCQ1 = Zurich Claudication Questionnaire Part 1; ZCQ2 = Zurich Claudication Questionnaire Part 2; ZCQ3 = Zurich Claudication Questionnaire Part 3; ODI = Oswestry Disability Index; PCS = Physical Composite Subscale; BVAS = Back Visual Analog Scale; LVAS = Leg Visual Analog Scale; PF = Physical Function; BP = Bodily Pain; RE = Role Emotional; MH = Mental Health; BMI = Body Mass Index.

Table 9

Overall Goodness of Fit and Chi-square Difference Test Results

Model		Goodness of fit parameters χ^2 (*df*)	RMSEA (90% conf int)	χ^2 DIFF (*df*)	*p*	Models Compared
1.0	Initial Model	1104.5 (284)	0.105 (0.098; 0.111)			None
1.1	Initial Model with ZCQ3 removed	931.1 (237)	0.106 (0.098; 0.113)	173.4 (47)	<0.0001	1.1 vs. 1.0
1.2F	Modified Model - Final	282.3 (114)	0.075 (0.068; 0.089)	648.8 (123)	<0.0001	1.2 vs. 1.1
2.1	Across occasion factor loadings - constrained equal	426.7 (128)	0.096 (0.087; 0.107)	144.4 (14)	<0.0001	2.1 vs. 1.2
2.2F	Factor loadings freed - ZCQ2, PF, BVAS, BP, & RE	302.5 (116)	0.078 (0.068; 0.089)	124.2 (12)	<0.0001	2.2 vs. 2.1
3.0	Model 2.2F plus Treatment group	297.5 (128)	0.071 (0.060; 0.081)			Not applicable
3.1	Model 2.2F plus all exogenous variables	370.3 (176)	0.065 (0.056; 0.074)			Not applicable

Note. F denotes final model for steps 1 and 2. RMSEA = root mean square error of approximation; χ^2DIFF = chi-square difference.

Model 1.0. First, I reviewed the theoretically justified structural equation model outlined in Chapter 3 (Figure 3). In this model, the observed variables ZCQ1, ZCQ2, ODI, PCS, BVAS, and LVAS were associated with PQoL. Variables ZCQ3, MH, and RE were associated with MQoL (Table 8, Model 1.0). Based on chi-square and RMSEA statistics, this model was not a good fit to the data (Table 9, Model 1.0). King-Kallimanis et al. (2011) highlighted that in longitudinal models the pattern of factor loadings must be consistent across all occasions. As Model 1.0 did not meet this criterion, I removed ZCQ3 from loading on MQoL at the 3- and 12-month occasions. ZCQ3 could not be added to the baseline timepoint as the instrument did not include ZCQ3 prior to an intervention.

Model 1.1. By testing the adjusted Model 1.1 using LISREL 9.1 software, I found the model did not fit the data (Table 9, Model 1.1) as the chi-square test was significant and RMSEA was greater than 0.08 (RMSEA = 0.108).

Model 1.2F. Using the pre-specified adjustments outlined in Chapter 3 Respecification Alternatives, I investigated a number of alternative models by substituting, replacing, and removing observed variables. Through this process, I tested an updated model that associated the observed variables of ZCQ2, BVAS, PF, and BP with PQoL; and RE and MH with MQoL (Table 9, Model 1.1). The fit was unsatisfactory (RMSEA = 0.106). Inspection of the modification indices suggested cross-loading of RE onto PQoL (including role emotional in both PQoL and MQoL assessments) and the addition of an error covariance between BVAS and PF (back pain and general physical function). I determined these associations to be theoretically sound. No other suggested

modifications could be supported by theory and so I did not include them. The changes produced a model with satisfactory fit (Table 9, Model 1.2F) where even though the chi-square test was significant, the RMSEA was less than 0.08 indicating the model had reasonable fit. The path diagram and output files are found in Appendix B. The final model was theoretically justified as condition-specific and general wellbeing QoL scores documenting observed function, pain, and mental from several instruments were combined. As the model fit was satisfactory and interpretation clear, I proceeded to the next step.

Model 2.1. To evaluate across occasion response shift, I constrained all factor loadings and intercepts to be equal across the three measurement occasions and created Model 2.1. The fit of this model was significantly worse when compared to Model 1.2F (χ^2 DIFF = 144.4, *df* = 14, p = <0.0001, $\alpha = 0.05$) indicating that response shift existed between measurement occasions (Table 9, Model 1.2F). This analysis supported the first stage of Hypothesis 1 testing. To comply with the hypothesis testing plan further evaluation of individual response shift components was required so I conducted the next stage of assessments.

Model 2.2F. To identify the magnitude and type of response shift at the 3- and 12-month occasions, I optimized the model by removing the equality constraints from each variable independently and comparing the fit of the resulting model. In each series, I freed the observed variable that was both significant and yielded the largest improvement. This process was repeated until no models in the next series showed significant improvement in the chi-square difference parameter. I created and assessed a

total of 30 models (Appendix C). To minimize family-wise error and reduce chance findings in this iterative testing, I determined significance using the Bonferroni adjusted level of significance $\alpha^* = \alpha_f/(n_z n_t)$ where α_f was the family-wise level of significance, n_z, was the number of factor loadings fixed at zero for a single measurement occasion, and n_t was the number of measurement occasions (King-Kallimanis, 2010).

Based on this iterative assessment, I identified that freeing RE, BP, ZCQ2, BVAS, and PF would improve model fit and confirmed that response shift existed in these variables across measurement occasions. Comparison of the final Model 2.2F to Model 2.1 indicated that Model 2.2F was improved (χ^2 DIFF = 124.2, *df* = 12, p = <0.0001, $\alpha = 0.05$). I performed this analysis to develop the model rather than to conduct hypothesis tests. I then continued my evaluation by assessing individual response shift components.

For detection of reconceptualization and using Model 2.2F, I assessed the factor-loading patterns between 3 and 12 months and identified no change (Table 10). Therefore, no reconceptualization between 3 months and 12 months was demonstrated.

Table 10

Standardized Factor Loadings for Model 2.2F

			PQoL			MQoL	
	ZCQ2	BVAS	PF	BP	RE	RE	MH
Baseline Factor Loadings	0.66	0.36	-0.64	-0.60	-0.19	0.61	0.93
3-month Factor Loadings	0.81	0.66	-0.77	-0.85	-0.19	0.43	0.97
12-month Factor Loadings	0.87	0.83	-0.83	-0.87	-0.42	0.36	1.03
Baseline and 3-month Factor Loadings Difference	**0.15**	**0.30**	**0.13**	**0.25**	0	**0.18**	0.04
Baseline and 12-months Factor Loadings Difference	**0.21**	**0.47**	**0.19**	**0.27**	**0.23**	**0.25**	**0.10**
3-month and 12-month Factor Loadings Difference	0.06	**0.17**	0.06	0.02	**0.23**	0.07	0.06

Note. Differences in factor loadings ≥ .10 are in boldface. PQoL = Physical Quality of Life; MQoL = Mental Quality of Life; ZCQ2 = Zurich Claudication Questionnaire Part 2; BVAS = Back Visual Analog Scale; PF = Physical Function; BP = Bodily Pain; RE = Role Emotional; MH = Mental Health.

Testing for reprioritization, I compared the magnitude of factor loadings between measurement occasions (Table 10). Reprioritization response shift was identified between baseline and both follow-up timepoints as multiple loadings differed by at least 0.10. All observed variables were impacted between baseline and 12 months and a majority of variables (71%) were impacted between baseline and 3 months. Since I did not identify significant response shift in MH in Stage 1, the borderline factor-loading difference in this variable at 12 months could be a chance finding.

For assessment of Hypothesis 1, I focused on the differences between the 3-month and 12-month timepoints and identified a change in factor-loading magnitude of at

least 0.10 for two variables—the BVAS standardized factor loadings differed by 0.17 and RE loadings on PQoL differed by 0.23 (Table 10). Therefore, the change in factor-loading magnitude was significant and indicated reprioritization response shift between 3 month and 12 months.

Based on the two-stage Hypothesis 1 analysis, I rejected the null hypothesis. In Stage 1, response shift was identified based on the comparison of Model 1.2F and Model 2.1. Specifically, the chi-square difference test was significant ($p = <0.0001$, $\alpha = 0.05$) denoting the models were not equivalent. In Stage 2, I identified reprioritization response shift between 3 months and 12 months based on a difference of at least 0.10 in the factor loading magnitude of two observed variables, BVAS and RE. Based on the rejection of the null hypothesis, I concluded that response shift at 12 months was different than response shift at 3 months. As no further testing of hypothesis 1 was required, I moved on to the analysis of hypothesis 2 and did not evaluate the model for recalibration response shift components.

Influence of Treatment Group on Response Shift

Hypothesis 2 testing required the addition of exogenous variables, specifically treatment group, to the final Step 2 model to assess the direct effects of patient and surgical characteristics. To support this analysis, I created and assessed two additional models, Model 3.0 and 3.1 (Table 8). Consistent with earlier analysis, model goodness of fit was assessed based on chi-square and RMSEA. Direct effects were assessed based on chi-square difference statistics using Bonferroni adjusted levels of significance.

Model 3.0. To create Model 3.0, I added the treatment group variable to Model 2.2F and allowed it to correlate with PQoL and MQoL. I fixed all direct effects on the observed variables to zero. The fit of Model 3.0 was satisfactory ($\chi^2 = 297.5$, $df = 128$, RMSEA = 0.071). Since the relationship between Model 3.0 and Model 2.2F was not hierarchical, the chi-square difference test was not a valid test statistic and was not assessed (Kline, 2011).

To test for direct effects, I created a series of models associating treatment group with each observed variable separately. The series contained six models and I assessed significance based on $\alpha^* = 0.0083$. In this series when I associated treatment group and MH, the LISREL software produced illogical results and the warning that the covariance matrix was not positively definitive. As the estimation methods employed by the LISREL software require a positive definite matrix, I determined that this result was uninterpretable and could not support the finding of a direct effect. The variable MH was removed from the analysis procedure.

I did not identify any direct effects based on treatment group as no unconstrained models met the criteria of a significant chi-square difference test (at adjusted significance levels) and parameter changes greater than 0.10. Based on these data, I did not reject the null hypothesis that subjects who received treatment A did not experience a difference in response shift from subjects who received treatment B between any measurement occasions. Since I detected no response shift bias based on treatment group, I further determined that there was no evidence to support that response shift influenced the direct comparison of QoL data between baseline and 12 months in the ISISS study.

Model 3.1. To provide additional information concerning the impact of exogenous variables on the study results, I added age, gender, levels treated, and BMI to Model 3.0 to create Model 3.1 (Table 8). Exogenous variables were allowed to correlate with latent variables while all direct effects on the observed variables were fixed to zero (Figure 4). While available, I did not include race as an exogenous variable as my review of the demographic data found the population to be almost homogenous at 95% white (Table 7).

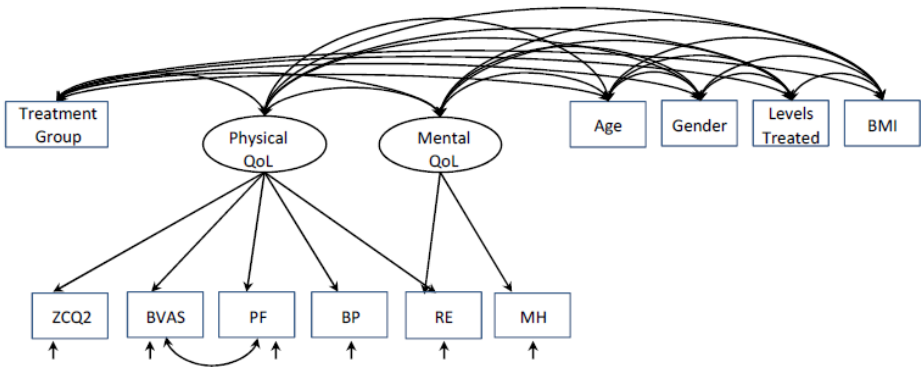


Figure 4. Graphical display of Model 3.1 showing first occasion variables only. QoL = quality of life; BMI = body mass index; ZCQ2 = Zurich claudication questionnaire part 2; BVAS = back visual analog scale; PF = physical function; BP = bodily pain; RE = role emotional; MH = mental health.

The fit of Model 3.1 was satisfactory ($\chi^2 = 370.3$, $df = 176$, RMSEA = 0.065). The path diagram and output files are found in Appendix D. In my review of the resulting covariance matrix (Table 11), I confirmed that treatment group had an extremely small covariance with PQoL and MQoL across all occasions. I interpreted this as not contradicting the previous finding that treatment group had no direct effect even when multiple exogenous factors were included in the model.

Table 11

Model 3.1 Covariances

	PQoL T1	MQoL T1	PQoL T2	MQoL T2	PQoL T3	MQoL T3
PQoL -T1	1					
MQoL -T1	-0.37	1				
PQoL –T2	0.46	-0.18	1			
MQoL –T2	-0.22	0.47	-.050	1		
PQoL –T3	0.47	-0.18	0.73	-0.43	1	
MQoL –T3	-0.18	0.42	-0.30	0.49	-0.47	1
Treatment group	-0.04	0.03	0.03	-0.02	-0.01	-0.01
Age	-0.03	1.72	1.71	0.73	1.80	0.10
Gender	0.13	-0.07	0.06	-0.04	0.04	-0.03
Levels treated	0.06	0.01	0.03	0.002	0.03	-0.01
BMI	1.10	-0.46	0.33	-0.53	-0.07	-0.09

Note. PQoL = Physical Quality of Life; MQoL = Mental Quality of Life; BMI = Body Mass Index.

To detect significant direct effects in this updated model, I repeated the analysis process outlined for Model 3.0 by created a series of iterative models. The series contained 30 models and significance was assessed based on $\alpha^* = 0.016$. As in the earlier Step 3 analysis, models created to assess MH for all exogenous variables were uninterpretable and removed from consideration for being freed. In the first series, no significant direct effects were associated with treatment group at any occasion. Significant direct effects of age and gender on RE were suggested but as the results from

this series supported no direct effect of treatment group, I did not follow-up with the additional testing to fully characterize the model's direct effects. I concluded that the analysis of adding additional exogenous factors to the model did not raise any questions concerning the retention of the null hypothesis for Hypothesis 2 and provided additional support for the previous findings.

Summary

I concluded that the answer to research question 1 was that back pain patients in the ISISS study did experience a difference in response shift between baseline and 3 months and between baseline and 12 months postintervention based on the rejection of the null hypothesis. I rejected the null hypothesis as a result of my detection of response shift in the overall model (Stage 1 test) in addition to the identification of reprioritization response shift between 3 months and 12 months (Stage 2 test). Specifically in Stage 1, the chi-square difference test between models 1.2F and 2.1 was significant ($p = <0.0001$, $\alpha = 0.05$) indicating that the models were not equivalent and overall response shift was detected. In Stage 2, my comparison of the magnitude of factor-loading values between occasions identified a difference greater than 0.10 between the 3 month and 12 month measurement occasions that I interpreted as reprioritization response shift between these two follow-ups.

For research question 2, I found there was insufficient evidence to conclude that response shift phenomenon impacted the clinical comparison of patient-reported outcomes between baseline and 12 months in the ISISS study based on the failure to reject the null hypothesis. I did not reject the null hypothesis that subjects who received

treatment A did not experience a difference in response shift from subjects who received treatment B between any measurement occasions as a result of my SEM analysis into direct effects using Model 3.1. Specifically, I identified no unconstrained models in my iterative analysis of treatment group that met the criteria of a significant chi-square difference test ($\alpha = 0.0083$) and parameter changes greater than 0.10. Therefore since a difference in response shift between the two treatment groups was not detected, I determined ISISS investigators could conduct direct clinical comparison of the treatment group results without accounting for differing response shift influences. In Chapter 5, I present further discussion of the results of this modeling study and implications for the integration of response shift analysis in the clinical interpretation of clinical data. I also explore how this research can result in positive social change and make recommendations for practice and future research.

Chapter 5: Discussion, Conclusions, and Recommendations

In this research, I investigated the impact of response shift phenomenon in a lumbar spinal stenosis population after receiving a surgical intervention. Using SEM techniques, I identified reprioritization at both 3 and 12 months with this response shift changing between the two follow-up timepoints. In my analysis of exogenous factor influence on response shift, I found that treatment group did not significantly impact the QoL reporting at 12 months. These findings add to the response shift body of knowledge by documenting that response shift was found in a spinal intervention population, reprioritization was different at different timepoints, and that direct comparisons between treatment groups can be made with no requirement to adjust for response shift. In this chapter, I expand on the interpretation of the findings, discuss other aspects of the research, and make recommendations for practice and future research.

Interpretation

In the evaluation of the clinical data, I identified response shift that differed between the 3- and 12-month follow-ups. The change in factor loadings between measurement occasions supported the finding that reprioritization response shift impacted BVAS (back pain as measured by the VAS pain scale) and RE (role limitations based on emotional issues). As reprioritization represents a change in the way a patient values specific aspects of his or her health, this change indicated that the back pain the patients experienced at 12 months had a greater influence on their perceived physical well-being than it did at 3 months. Likewise, I found that the correlation between RE and the patient's PQoL scores increased between the 3- and 12-month follow-ups. The

correlation was negative because, unlike the other scales that make up PQoL, increased SF-12 measures indicated improved patient status. Reprioritization change in back pain could indicate that patients at 3 months may not have considered themselves fully healed from their surgery and so they associated less of their physical wellbeing to their back pain. It is also possible that the general reduction in pain and the fact that some patients may still have been medicated for surgical pain could also influence the relationship between back pain and overall QoL. Conversely, most patients at 12 months had accepted that the level of pain they were experiencing had stabilized. This realization could influence them to assign more importance to the pain they were experiencing when scoring their QoL.

In a similar way, physical limitations due to emotional conditions such as depression and anxiety may be included in the patient's general QoL scores differently at 3 and 12 months. Because patients were completing the instruments on a regular basis and, in general, their physical status was improving, it is also possible that the respondents were better able to differentiate between emotional and physical causes for limitations in their day-to-day lives. These factors could also support reprioritization response shift. It is interesting to note that RE retained similar importance, no reprioritization, when predicting MQoL between these two follow-ups. These easily understood explanations can be used to support the reasoning that the response shift identified in the data was not a chance finding.

In the evaluation of patient and surgical characteristics on response shift, the addition of exogenous variables in both Models 3.0 and 3.1 improved the model fit to the

data (Table 9). However, unlike in drug trials where both patient and investigator can be blinded to treatment arm, many device trials, especially those that involve surgical implants, cannot keep this information from the subject. Receiving the experimental or control treatment may influence a patient's expectations for recovery and symptom resolution, as the potential benefits of the new treatment are covered extensively in the informed consent phase of a clinical trial. Disappointment in treatment assignment, an expectation of faster or miraculous healing if the experimental treatment was received, or other differences in expectations could influence QoL reporting and thereby bias a direct comparison of the outcome measures. Finding no direct effects of treatment group on response shift supported the direct comparison of the two study arms. While randomized clinical trial designs implemented by experienced research staff should reduce treatment group bias significantly, having an additional tool for confirmation could increase confidence in the clinical conclusions made by investigators. In this research, the identification of a potential direct effect associated with age and gender could be used to better understand specific subpopulations, direct future research, and prevent inaccurate ad hoc conclusions concerning the results.

Discussion

The response shift theoretical model includes the requirement for a catalyst, defined as a change in an individual's health state regardless of source or direction (Sprangers & Schwartz, 1999). Therefore, the identification of a significant difference in response shift between 3 and 12 months supported time, in addition to the intervention, as a catalyst event in this spine intervention population. This finding was consistent with a

previous orthopedic intervention study that identified response shift between 6 and 12 months after surgery and found response shift confounded the accurate measurement of patient recovery (Razmjou et al., 2009). However, when evaluating response shift in other medical conditions, researchers have documented varied results. In multiple sclerosis patients followed for 18 months, King-Kallimanis et al. (2011) identified only a small response shift. As the population was selected based on an expectation of large response shift, this finding was unexpected and the researchers identified the lack of a catalyst event as one of the reasons for the findings. In a study of an HIV/AIDS population that covered 2 years, King-Kallimanis et al. (2010) found minimal response shift. The researchers again highlighted the lack of a specific catalyst event such as a new diagnosis or an intervention as a study limitation. Conversely, in a longitudinal study of older men who both had strokes and were stroke free, reprioritization response shift was identified in all groups (Barclay & Tate, 2014). Covering 4 years, these results could indicate that time did serve as a sufficient catalyst event in the stroke-free older male population. The limited body of knowledge related to response shift and catalyst events does not provide any actionable results. However, whether time serves as a trigger for response shift could have significant implications for medical research as, if confirmed, this finding could invalidate accepted conclusions from previous longitudinal trials. Increased understanding of the impact in different acute and chronic conditions could also require significant changes in accepted clinical trial design and expectations.

While theoretically complex, I was able to assess response shift in patient-reported outcome measurements effectively from a comparative clinical trial by applying

SEM techniques. The process was fairly time consuming but the final results were interpretable and credible. By tailoring the process to focus on aspects relevant to comparative clinical trials, I was able to simplify the technique to a small degree. The primary challenge I had implementing SEM was in identifying an adequate starting measurement model because no other researchers had validated a SEM QoL model for a lumbar spinal stenosis surgical population. One early decision I had to make was whether to include only one or multiple instruments in the model. Literature did not provide any recommendations as both single and multiple instrument models had been used successfully (Ahmed et al., 2009; Barclay & Tate, 2014; Barclay-Goddard et al., 2009; King-Kallimanis et al., 2010). I elected to incorporate multiple instruments because five instruments were used in the ISISS study to support the analysis of primary and secondary endpoints. In addition, it was not feasible to use only the primary assessment instrument, the ZCQ, as the limited number of parameters violated model identification requirements. I did try to include at least one ZCQ variable in the final model as this instrument was critical to the primary treatment comparison and it was the only disease specific measurement tool. Because researchers had identified that even in the same studies, spine patients reported diverse and conflicting results when different instruments were used (Copay et al., 2010; Schwartz & Finkelstein, 2009), the difficulty in identifying a starting model was not unexpected. Another decision I made in the SEM process was whether to require reasonable fit (RMSEA $\leq$ 0.08) or close fit (RMSEA $\leq$ 0.05) goodness of fit criteria. I selected reasonable fit as sufficient because the purpose of my research was to assess response shift in the context of interpreting comparative

clinical data and not to determine cause and effect relationships or to quantify true change. Other researchers had made the same decision (Ahmed et al., 2009; Barclay & Tate, 2014).

In the health care literature, I found a general consensus on the importance of assessing response shift when QoL endpoints were used to support clinical efficacy decisions (Ahmed et al., 2005; Barclay & Tate, 2014; Hamidou et al., 2011; King-Kallimanis et al., 2012; Razmjou et al., 2009; Schwartz et al., 2006). However, there was significant variability in the ability of different methods to identify the same response shift. In direct comparisons, different methods yielded discrepant results in type of response shift, effect size, clinical significance, or a combination of these (Ahmed et al., 2005; Nagl & Farin, 2012; Visser et al., 2005). Therefore, with no validated or standard methodology for determining true change, how should investigators address response shift when it is identified? How can the potential for response shift bias be minimized as part of the clinical study design? In some cases, an evaluation of response shift will confirm that this phenomenon does not invalidate a direct comparison. In this research, the finding that treatment group did not influence the data at 12 months provided an example of this and the ISSIS investigator would be correct in implementing standard comparative analysis. However, investigators cannot count on the absence of significant response shift and therefore methods to address this phenomenon should be included in the initial study design. Techniques used to address other potential biases could be adapted such as including prespecified decision points and significance levels in study procedures and analysis plans. To do this, investigators would need to consider and

document the potential relationships between the types of response shift and research questions, study design, data structure, and specific QoL instruments. This could be a monumental task since relevant parameters increase exponentially when complex relationships are modeled. An additional complication with this approach would be that when response shift was identified, the investigator would need to report specifically how the information was integrated into the clinical interpretation. At a minimum, this should include the impact of recalibration, reprioritization, and reconceptualization.

Alternatively, given the lack of accepted response shift adjustment information, investigators could select a clinical study design to minimize the potential impact of response shift. Eliminating the requirement to adjust QoL measures for response shift could be accomplished by using an equivalency or non-inferiority hypothesis in place of a direct comparison or superiority analysis. However, these equivalency designs could impose significant restrictions since a comparison treatment that was both equivalent and ethically valid would be required. Additionally, treatments that were superior to existing options, a typical goal of new interventions, often would not meet equivalency criteria.

Limitations of the study

During analysis, I did not identify any additional limitations to this research. However, the previously identified limitations remained. First, the study design was based on secondary analysis of previously collected experimental clinical trial data and as such the design did not drive data collection. However, sufficient data to support the SEM methodology was available so this restriction did not have a negative impact on the study. Data screening and cleaning procedures did not identify any significant data issues.

The mitigation of the source data being collected in accordance with good clinical practices that included electronic data entry with built-in edit checks and regular monitoring by the sponsor reduced concerns associated with the fact that the data was collected prior to final quality checks.

The potentially subjective nature of advanced modeling procedures also did not introduce any unexpected research limitations. The evaluation methodology and goodness of fit criteria outlined prior to the analysis were implemented without adjustment. By prespecifying alternative models and variables, the potential for significant bias to be introduced into model respecification was minimized. Additionally, all model adjustments were able to be theoretically justified.

Recommendations

Instead of being isolated in QoL research, incorporating response shift assessment into clinical study design would benefit all clinical investigators that use patient-reported outcomes as endpoints. Due to the potential for response shift phenomenon to invalidate efficacy and treatment effect conclusions, this recommendation applies across research disciplines. An initial step would be to continue to communicate to researchers the potential confounding effect of response shift on standard clinical trial design and to provide tools to detect response shift in specific datasets. Publication in peer-reviewed journals outside of the QoL discipline would be beneficial. These articles would also serve to educate regulatory agencies, reimbursement professionals, and medical societies on the importance of considering response shift in clinical data interpretation. Additionally, as the body of knowledge concerning response shift adjustment is so

limited with no validated methodologies available, clinical study designers should give preference to equivalency and non-inferiority study designs over direct comparisons and superiority hypotheses. This runs counter to the current emphasis by physicians and governments on comparative effectiveness research. Equivalence designs often do not identify a clear treatment preference and the resulting increase in treatment options could increase rather than decrease health care costs. However, since the trade-off would be between investigators providing accurate clinical conclusions and the preference of regulatory and reimbursement professionals, patients would be best served by study designs that minimize response shift bias until investigators can accurately and consistently support true change comparisons.

I also advocate additional research into the entire response shift discipline. Despite being explored for the past decade, response shift phenomenon research is still in an early stage. Research into the catalysts, antecedents, and mechanisms of response shift and how to identify and measure these variables would support the theory and foundation of this phenomenon. To expand the body of knowledge on response shift methodologies, comparisons of approaches, the validation of methods to accurately adjust for response shift, and the application of new and existing methods to varied QoL instruments and medical conditions are needed. Of particular value would be research that directly compares multiple techniques for response shift identification and quantification in a population where response shift had been previously characterized. Studies to expand SEM and other statistical methods that use secondary data analysis, the development of consistent reporting standards, and validated clinical practice guidelines would support

the practical integration of response shift into research disciplines outside of QoL. My identification of reprioritization in spine intervention subjects demonstrates that response shift is important to the accurate interpretation of patient-reported QoL outcome data so I recommend further research into all aspects of the phenomenon.

Implications for Social Change

The findings of this research provided support for including response shift evaluation into the clinical interpretation of QoL data to prevent false or inaccurate conclusions. The understanding that time can serve as a catalyst for response shift in a spine intervention population will support improved QoL adjustments for individuals living with chronic spine conditions. This study also provided clinical investigators outside of the QoL discipline with a practical methodology for evaluating the clinical significance of response shift on previously collected data. These additions to the response shift body of knowledge support an increased understanding of how this phenomenon can confound or invalidate accepted clinical study data interpretation. This insight can support physicians in coming to accurate clinical conclusions and enhance clinical decision-making. The incorporation of response shift evaluation into clinical study design will support the accurate interpretation of clinical trial data and translate into improved health outcomes for patients worldwide.

Conclusion

Using this research, I investigated the impact of response shift on the clinical interpretation of comparative data in a spine intervention clinical study. Response shift phenomenon can interfere with clinical data interpretation when patients adjust the

framework they use to score their QoL at different occasions. While response shift has been identified in health care data, the impact in interventional spine studies had not been studied. Using SEM, which could be applied to previously collected data, I identified a significant difference in response shift between the 3-month and 12-month follow-ups; a finding that could invalidate conclusions based on a direct comparison of QoL scores at these timepoints. However, since no difference in response shift associated with treatment group at 12 months was identified, a direct comparison of the results was appropriate. These findings should be considered when assessing ISISS secondary endpoints but do not impact the primary study analysis as intervention success by patient served as the primary outcome variable. Success was determined based on a combination of individual clinically significant improvement, postintervention treatments, and adverse events.

When treatment effects are being quantified based on patient-reported measures, clinical investigators should incorporate the assessment of response shift into the interpretation of the clinical data. Failure to investigate this phenomenon could result in inaccurate conclusions due to under- or overestimating treatment effects. SEM can be used to perform this assessment and to further explore the impact of response shift on clinical study results and conclusions.

References

Ahmed, S., Bourbeau, J., Maltais, F., & Mansour, A. (2009). The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. *Journal of Clinical Epidemiology*, *62*(11), 1165–1172. doi:10.1016/j.jclinepi.2009.03.015

Ahmed, S., Mayo, N.E., Wood-Dauphinee. S., Hanley, J.A. & Cohen, S.R. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *Journal of Clinical Epidemiology*, *58*, 1125–1133. doi:10.1016/j.jclinepi.2005.03.003

Anderson, P. A., Carreon, L. Y., & Glassman, S. D. (2009). Response shift phenomenon. Does this apply to spine outcomes research? *The Spine Journal*, *9*(12), 1037–1038. doi:10.1016/j.spinee.2009.08.449

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423. doi:10.1037/0033-2909.103.3.411

Barclay, R., & Tate, R. B. (2014). Response shift recalibration and reprioritization in health-related quality of life was identified prospectively in older men with and without stroke. *Journal of Clinical Epidemiology*, *67*(5), 500–507. doi:10.1016/j.jclinepi.2013.12.003

Barclay-Goddard, R., Epstein, J.D., & Mayo, N.E. (2009). Response shift: A brief overview and proposed research priorities. *Quality of Life Research, 18*, 335-346. doi:10.1007/s11136-009-9450-x

Barclay-Goddard, R., Lix, L.M., Tate, R., Weinberg, L. & Mayo, N.E. (2009). Response shift was identified over multiple occasions with a structural equation modeling framework. *Journal of Clinical Epidemiology*, *62*(11), 1181–1188. doi:10.1016/j.jclinepi.2009.03.014

Bernhard, J., Hürny, C., Maibach, R., Herrmann, R., & Laffer, U. (1999). Quality of life as subjective experience: Reframing of perception in patients with colon cancer undergoing radical resection with or without adjuvant chemotherapy. *Annals of Oncology*, *10*(7), 775–782. Retrieved from http://annonc.oxfordjournals.org/

Blair, H., Wilson, L., Gouick, J., & Gentleman, D. (2010). Individualized vs. global assessments of quality of life after head injury and their susceptibility to response shift. *Brain Injury*, *24*(6), 833–843. doi:10.3109/02699051003789203

Calvert, M., Blazeby, J., Altman, D.G., Revicki, D. A., Mother, D., & Brundage, M.D. for the CONSORT PRO Group (2013). Reporting of patient-reported outcomes in randomized trials: The consort pro extension. *JAMA*, *309*(8), 814–822. doi:10.1001/jama.2013.879

Carver, C. S., & Scheier, M. F. (2000). Scaling back goals and recalibration of the affect system are processes in normal adaptive self-regulation: Understanding 'response shift' phenomena. *Social Science & Medicine*, *50*(12), 1715-1722. doi:10.1016/S0277-9536(99)00412-8

Copay, A. G., Glassman, S. D., Subach, B. R., Berven, S., Schuler, T. C., & Carreon, L. Y. (2008). Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical

Outcomes Study questionnaire Short Form 36, and Pain Scales. *The Spine Journal*, *8*(6), 968-974. doi:10.1016/j.spinee.2007.11.006

Copay, A. G., Martin, M. M., Subach, B. R., Carreon, L. Y., Glassman, S. D., Schuler, T. C., & Berven, S. (2010). Assessment of spine surgery outcomes: inconsistency of change amongst outcome measurements. *The Spine Journal*, *10*(4), 291-296. doi:10.1016/j.spinee.2009.12.027

Don, A. S., & Carragee, E. (2008). A brief overview of evidence-informed management of chronic low back pain with surgery. *The Spine Journal*, *8*(1), 258–265. doi:10.1016/j.spinee.2007.10.027

Donaldson, G. W. (2005). Structural Equation Models for Quality of Life Response Shifts: Promises and pitfalls. *Quality of Life Research*, *14*(10), 2345–2351. doi:10.1007/s11136-005-3977-2

Fairbank, J. C., & Pynsent, P. B. (2000). The Oswestry Disability Index. *Spine*, *25*(22), 2940. doi:10.1097/00007632-200011150-00017

Fairbank, J. C., Couper, J., Davies, J. B., & O'Brien, J. P. (1980). The Oswestry Low Back Pain Disability Questionnaire. *Physiotherapy*, *66*(8), 271-273. Retrieved from http://pt.unlv.edu/ebpt/tests/

Finkelstein, J. A., Razmjou, H., & Schwartz, C. E. (2009). Response shift and outcome assessment in orthopedic surgery: Is there a difference between complete and partial treatment? *Journal of Clinical Epidemiology*, *62*(11), 1189-1190. doi:10.1016/j.jclinepi.2009.03.022

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and

persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science*, *12*(2), 133–157. doi:10.1177/002188637601200201

Hamidou, Z., Dabakuyo, T. S., & Bonnetain, F. (2011). Impact of response shift on longitudinal quality-of-life assessment in cancer clinical trials. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(5), 549–59. doi:10.1586/erp.11.57

Haro, H., Maekawa, S., & Hamada, Y. (2008). Prospective analysis of clinical evaluation and self-assessment by patients after decompression surgery for degenerative lumbar canal stenosis. *The Spine Journal*, *8*(2), 380–384. doi:10.1016/j.spinee.2007.01.010

Houweling, T.A.W. (2010). Reporting improvement from patient-reported outcome measures: A review. *Clinical Chiropractic*, 13, 15-22. doi:10.1016/j.clch.2009.12.003

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144–150. doi:10.1037/0021-9010.64.2.144

Jansen, S. J. T., Stiggelbout, A. M., Nooij, M. A., Noordijk, E. M., & Kievit, J. (2000). Response shift in quality of life measurement in early-stage breast cancer patients undergoing radiotherapy. *Quality of Life Research*, *9*(6), 603-615. doi:10.1023/A:1008928617014

Jenkinson, C., Layte, R., Jenkinson, D., Lawrence, K., Petersen, S., Paice, C., &

Stradling, J. (1997). A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? *Journal of Public Health*, *19*(2), 179-186. doi:10.1093/oxfordjournals.pubmed.a024606

King-Kallimanis, B. L., ter Hoeven, C. L., de Haes, H. C., Smets, E. M., Koning, C. C. E., & Oort, F. J. (2012). Assessing measurement invariance of a health-related quality-of-life questionnaire in radiotherapy patients. *Quality of Life Research*, *21*(10), 1745–1753. doi:10.1007/s11136-011-0094-2

King-Kallimanis, B. L., Oort, F. J., & Garst, G. J. A. (2010). Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *AStA Advances in Statistical Analysis*, *94*(2), 139–156. doi:10.1007/s10182-010-0129-y

King-Kallimanis, B.L., Oort, F.J., Nolte, S., Schwartz, C., & Sprangers, M. (2011). Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Quality of Life Research*, *20*(10), 1527–1540. doi:10.1007/s11136-010-9844-9

King-Kallimanis, B. L., Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2009). Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, *62*(11), 1157–1164. doi:10.1016/j.jclinepi.2009.04.004

Kline, R.B. (2011). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Kvam, A. K., Wisløff, F., & Fayers, P. M. (2010). Minimal important differences and

response shift in health-related quality of life; a longitudinal study in patients with multiple myeloma. *Health And Quality Of Life Outcomes*, *8*, 79-87. doi:10.1186/1477-7525-8-79

Lei, P. W., & Wu, Q. (2007). Introduction to Structural Equation Modeling: Issues and Practical Considerations. *Educational Measurement: Issues and Practice*, *26*(3), 33–43. doi:10.1111/j.1745-3992.2007.00099.x

Li, Y., & Rapkin, B. (2009). Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of Clinical Epidemiology*, *62*(11), 1138–1147. doi:10.1016/j.jclinepi.2009.03.021

Litcher-Kelly, L., Martino, S. A., Broderick, J. E., & Stone, A. A. (2007). A systematic review of measures used to assess chronic musculoskeletal pain in clinical and randomized controlled clinical trials. *The journal of pain : official journal of the American Pain Society*, *8*(12), 906–913. doi:10.1016/j.jpain.2007.06.009

Loguidice, V., Bini, W., Shabat, S., Miller, L. E., & Block, J. E. (2011). Rationale, design and clinical performance of the Superion® interspinous spacer: A minimally invasive implant for treatment of lumbar spinal stenosis. *Expert Review of Medical Devices, 8*(4), 419-426. doi:10.1586/erd.11.24

Luo, X., George, M. L., Kakouras, I., Edwards, C. L., Pietrobon, R., Richardson, W., & Hey, L. (2003). Reliability, validity, and responsiveness of the short form 12-item survey (SF-12) in patients with back pain. *Spine*, *28*(15), 1739-1745. doi:10.1097/01.BRS.0000083169.58671.96

Mayo, N.E., Scott, S.C., Dendukuri, N., Ahmed, S. & Wood-Dauphinee, S. (2008). Identifying response shift statistically at the individual level. *Quality of Life Research*, *17*, 627–639. doi:10.1007/s11136-008-9329-2

McPhail, S., & Haines, T. (2010a). The response shift phenomenon in clinical trials. *Journal of Clinical Research Best Practices*, *6*(2), 1-8.

McPhail, S. & Haines, T. (2010b). Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health and Quality of Life Outcomes*, 8, 1-9. doi:10.1186/1477-7525-8-65

Moore, A., Moore, O., McQuay, H., & Gavaghan, D. (1997). Deriving dichotomous outcome measures from continuous data in randomised controlled trials of analgesics: use of pain intensity and visual analogue scales. *Pain*, *69*(3), 311-315. doi:10.1016/S0304-3959(96)03306-4

Nagl, M., & Farin, E. (2012). Response shift in quality of life assessment in patients with chronic back pain and chronic ischaemic heart disease. *Disability & Rehabilitation*, *34*(8), 671–680. doi:10.3109/09638288.2011.619616

Nolte, S., Elsworth, G.R., Sinclair, A.J., & Osborne, R.H. (2009). Tests of measurement invariance failed to support the application of the "then-test". *Journal of Clinical Epidemiology, 62,* 1173-1180. doi:10.1016/j.jclinepi.2009.01.021

Olaogun, M.O.B., Adedoyin, R.A., Ikem, I.C., & Anifaloba, O.R. (2004). Reliability of rating low back pain with a visual analogue scale and a sematic differential scale. *Physiotherapy Theory and Practice, 20,* 135-142. doi:10.1980/09593980490453048

Oort, F. J. (2005a). Types of change in self-report data, In N. Andes, J. Blasius, C. Ven Dijkum, H. Ganzeboom, & H. Kleijer (Eds.), *Recent Developments and Applications in Social Research Methodology; Proceedings of the Sixth International Conference on Social Science Methodology.* Amsterdam, The Netherlands: TT-Publikaties.

Oort, F. (2005b). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*(3), 587–598. doi:10.1007/s11136-004-0830-y

Oort, F.J., Visser, M., & Sprangers, M. (2005). An application of structural equation modeling to detect response shift and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, 14, 599–609. doi:10.1007/s11136-004-0831-x

Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, *62*(11), 1126–1137. doi:10.1016/j.jclinepi.2009.03.013

Ormerod, R. J. (2010). Justifying the methods of OR. *The Journal of the Operational Research Society*, *61*(12), 1694–1708. doi:10.1057/jors.2009.147

Osborne, R.H., Hawkins, M., & Sprangers, M. A. G. (2006). Change of perspective: A measureable and desired outcome of chronic disease self-management intervention programs that violates the premise of preintervention/ postintervention assessment. *Arthritis & Rheumatism, 55*, 458-465.

doi:10.1002/art.21982

Pratt, R.K., Fairbank, J.C., & Virr, A. (2002). The reliability of the shuttle walking test, the Swiss spinal stenosis questionnaire, the Oxford spinal stenosis score, and the Oswestry disability index in the assessment of patients with lumbar spinal stenosis. *Spine*, 27(1), 84-91. doi:10.1097/00007632-200201010-0002

Quality of life. (n.d.). In *Collins English Dictionary*. Retrieved from http://www.collinsdictionary.com/dictionary/english/quality-of-life

Rapkin, B. D., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health & Quality of Life Outcomes*, *2*, 14–12. doi:10.1186/1477-7525-2-14

Razmjou, H., Schwartz, C. E., Yee, A., & Finkelstein, J. A. (2009). Traditional assessment of health outcome following total knee arthroplasty was confounded by response shift phenomenon. *Journal of Clinical Epidemiology*, *62*(1), 91–96. doi:10.1016/j.jclinepi.2008.08.004

Reeve, B. B. (2010). An opportunity to refine our understanding of "response shift" and to educate researchers on designing quality research studies: response to Ubel, Peeters, and Smith. *Quality of Life Research*, *19*(4), 473–475. doi:10.1007/s11136-010-9612-x

Resnik, L., & Dobrzykowski, E. (2003). Guide to outcomes measurement for patients with low back pain syndromes. *The Journal of Orthopaedic and Sports Physical Therapy*, *33*(6), 307. Retrieved from europepmc.org

Ring, L., Hofer, S., Heuston, F., Harris, D., & O'Boyle, C. A. (2005). Response shift

masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. *Health and Quality of Life Outcomes*, *3*, 55. doi:10.1186/1477-7525-3-55

Schwartz, C.E., Andresen, E.M., Nosek, M.A., Krahn, G.L., & RRTC Expert Panel on Health Status Measurement. (2007). Response shift theory: Important implication for measuring quality of life in people with disability. *Archives of Physical Medicine and Rehabilitation, 88*, 1561–1572. doi:10.1016/j.apmr.2006.12.032

Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A. G., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Quality of Life Research*, *15*(9), 1533–1550. doi:10.1007/s11136-006-0025-9

Schwartz, C. E., & Finkelstein, J. A. (2009). Understanding inconsistencies in patient-reported outcomes after spine treatment: Response shift phenomena. *The Spine Journal*, *9*(12), 1039-1045. doi:10.1016/j.spinee.2009.05.010

Schwartz, C. E., & Rapkin, B. D. (2004). Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health & Quality of Life Outcomes*, *2*, 16–11. doi:10.1186/1477-7525-2-16

Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine*, *48*(11), 1531-1548. doi:10.1016/S0277-9536(99)00047-7

Schwartz, C. E., Feinberg, R. G., Jilinskaia, E., & Applegate, J. C. (1999). An evaluation of a psychosocial intervention for survivors of childhood cancer: paradoxical

effects of response shift over time. *Psycho-Oncology*, *8*(4), 344–354. doi:10.1002/(SICI)1099-1611(199907/08)8:4<344::AID-PON399>3.0.CO;2-T

Schwartz, C., Sprangers, M., & Fayers, P. (2005). Response shift: you know it's there but how do you capture it? Challenges for the next phase of research. In P. Fayers, & R. D. Hays (Eds.), *Assessing quality of life in clinical trials. Methods and Practice* (pp. 275-290), Oxford, UK: Oxford University Press.

Schwartz, C., Sprangers, M., Oort, F., Ahmed, S., Bode, R., Li, Y., & Vollmer, T. (2011). Response shift in patients with multiple sclerosis: an application of three statistical techniques. *Quality of Life Research*, *20*(10), 1561–1572. doi:10.1007/s11136-011-0056-8

Scientific Software International, Inc. (2013). LISREL (Version 9.1) [Structural equation modeling software].

Sprangers, M. A. ., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science & Medicine*, *48*(11), 1507–1515. doi:10.1016/S0277-9536(99)00045-3

Sprangers, M. A. G., & Schwartz, C. E. (2010). Do not throw out the baby with the bath water: build on current approaches to realize conceptual clarity. Response to Ubel, Peeters, and Smith. *Quality of Life Research*, *19*(4), 477–479. doi:10.1007/s11136-010-9611-y

Stucki, G., Daltroy, L., Liang, M.H., Fossel, A.H., & Katz, J.(1995). Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *Journal of Clinical Epidemiology, 48*, 1369-

1378. doi:10.1016/0895-4356(95)00054-2

Stucki, G., Daltroy, L., Liang, M.H., Lipson, S., Fossel, A.H., & Katz, J.(1996). Measurement properties of a self-administered outcome measure in lumbar spinal stenosis. *Spine, 21*, 796-803. doi:10.1097/00007632-199604010-00004

Ubel, P. A., Peeters, Y., & Smith, D. (2010). Abandoning the language of "response shift": A plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. *Quality of Life Research*, *19*(4), 465-471. doi:10.1007/s11136-010-9592-x

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*(1), 35–50. doi:10.1207/s15327752jpa8701_03

VertiFlex, Incorporated. (2013, September 1). Investigating Superion™ in Spinal Stenosis [ISISS] (NLM Identifier: NCT00692276). Retrieved from ClinicalTrials.gov, National Library of Medicine website: http://clinicaltrials.gov/show/NCT00692276

Visser, M. R., Oort, F. J., & Sprangers, M. A. (2005). Methods to detect response shift in quality of life data: a convergent validity study. *Quality of Life Research*, *14*(3), 629-639. doi:10.1007/s11136-004-2577-x

Walsh, T. L., Hanscom, B., Lurie, J. D., & Weinstein, J. N. (2003). Is a condition-specific instrument for patients with low back pain/leg symptoms really necessary?: the responsiveness of the Oswestry Disability Index, MODEMS, and the SF-36. *Spine*, *28*(6), 607–615. doi:10.1097/01.BRS.0000050654.97387.DF

Ware Jr, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical care*, *34*(3), 220-233. doi:10.1097/00005650-199603000-00003

Wilson, I.B. (1999). Clinical understanding and clinical implications of response shift. *Social Science & Medicine*, *48*(11), 1577–1588. doi:10.1016/S0277-9536(99)00050-7

Yardley, L. & Dibb, B. (2007). Assessing subjective change in chronic illness: An examination of response shift in health-related and goal-oriented subjective status. *Psychology and Health*, 22, 813-828. doi:10.1080/14768320601124808

Appendix A: Dataset E-mail

From: Stephen Reitzler
To: SDC
Cc: Robin Carlson
Sent: Monday, June 4, 2012 4:41 PM
Subject: FW: VertiFlex ISISS Trial - WIRB Protocol #20080548 - Attention: WIRB Regulatory Department

Katie and Tammy, the process of getting some authorization to release the redacted clinical data to Robin Carlson has been a long and drawn out one, I'm sorry to say. Our own attorneys suggested that to remain compliant with HIPAA regulations, we would need a waiver from our IRB – even though virtually every identifier of patients, clinical sites, investigators, and even device types, have been removed. With much shaking of the trees, we finally got a response. As they indicate below (see e-mail from Viveca Burnette to Leslie Zaccari), and for the reasons shown, WIRB does *not* believe a waiver is required to release the redacted data to Robin. Since they are the "authorities" in this matter, I believe we are safe in giving Robin what she's asked for, albeit quite belatedly. When you have the latest data cut finalized, please forward it to Robin in the same format as you prepared *Vertiflex Research Dataset 27MAR2012.xls*.

Thanks very much for being patient, and for helping Robin out with this "little" project.

From: VertiFlex
Sent: Monday, June 04, 2012 4:32 PM
To: Stephen Reitzler
Subject: FW: VertiFlex ISISS Trial - WIRB Protocol #20080548 - Attention: WIRB Regulatory Department

FYI- from WIRB with regard to Robin's data request

From: Viveca Burnette
Sent: Monday, June 04, 2012 4:27 PM
To: VertiFlex
Subject: RE: VertiFlex ISISS Trial - WIRB Protocol #20080548 - Attention: WIRB Regulatory Department

Thank you for your inquiry to WIRB and for speaking with me this morning. After further discussing internally, I have confirmed that you do not need to request a full waiver under HIPAA for this specific circumstance for the following reasons:

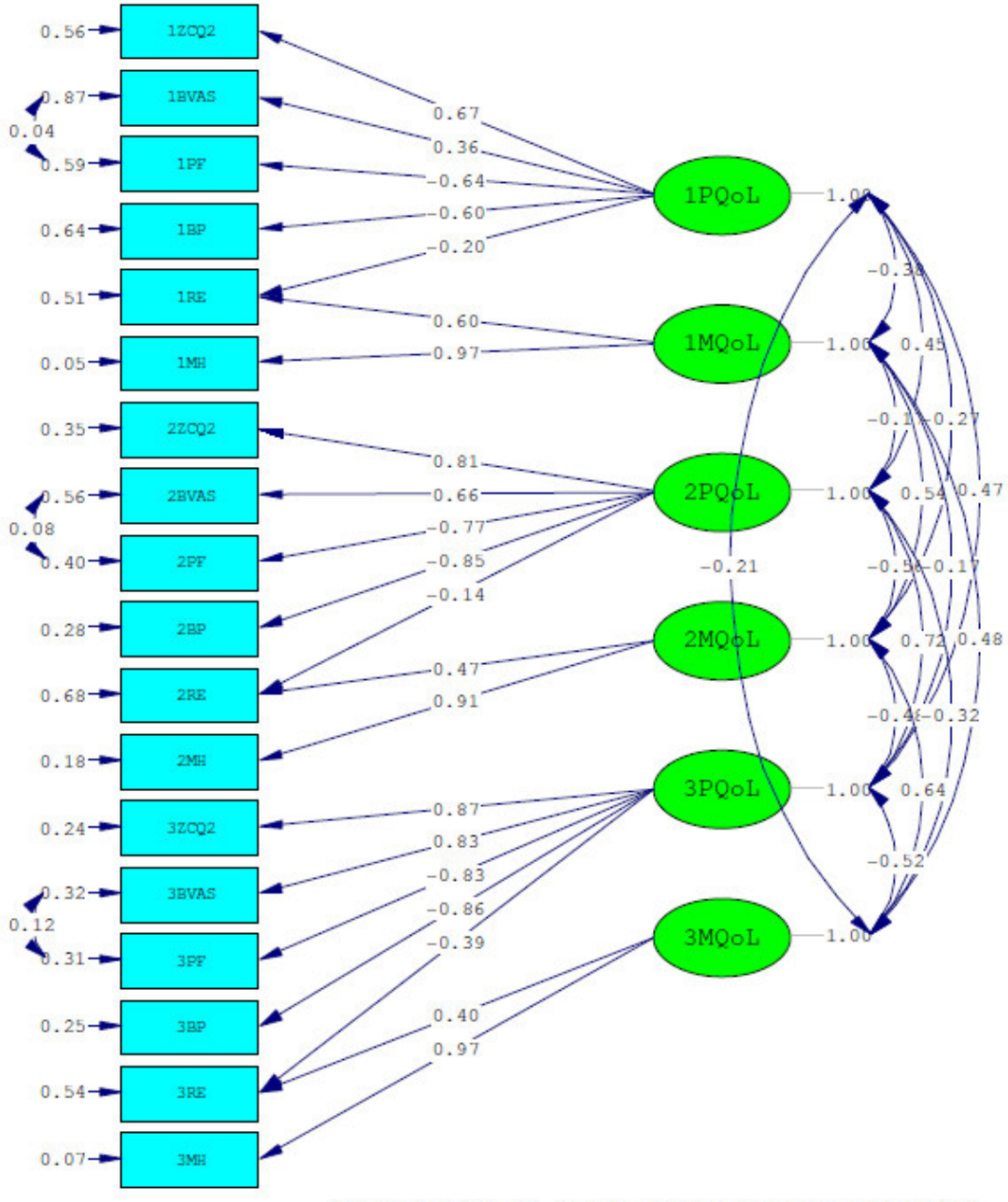
1. As VertiFlex (the sponsor) is a medical device company, it is not a covered entity as described in the HIPAA regulations;
2. Per your description of the information that will be provided to your former employee for use on her Ph.D. thesis paper, there is no information that will be originating from your organization that will include any subject identifiers.

Please let me know if I can provide any further assistance or if you need additional information.

Kind regards,
Viveca

Viveca Burnette, BA | Regulatory Analyst
Office: (360) 252-2880 | **Fax:** (360) 252-2525
vburnette@wirb.com | www.wirb.com
WESTERN INSTITUTIONAL REVIEW BOARD®

Appendix B: Model 1.2F LISREL Standardized Path Diagram and Output File



```
                                DATE:  9/25/2014
                                TIME:  7:53


                             L I S R E L  9.10 (32 Bit)

                                       BY

                             Karl G. Jöreskog & Dag Sörbom


                        This program is published exclusively by
                        Scientific Software International, Inc.
                            http://www.ssicentral.com

             Copyright by Scientific Software International, Inc., 1981-2012
               Use of this program is subject to the terms specified in the
                           Universal Copyright Convention.

The following lines were read from file C:\Users\rsc\Desktop\Research\Model 1.2.spj:

Model 1.2 - respectified Model
Raw Data from file 'C:\Users\rsc\Desktop\Research\RSData_Longitudinal.LSF'

                 ----------------------------
                  EM Algorithm for missing Data:
                 ----------------------------

               Number of different missing-value patterns=        5
               Effective sample size:      263

               Convergence of EM-algorithm in     4 iterations
               -2 Ln(L) =    29653.61988
               Percentage missing values=   1.84

          Note:
            The Covariances and/or Means to be analyzed are estimated
            by the EM procedure and are only used to obtain starting
            values for the FIML procedure

Sample Size = 263
Latent Variables  1PQoL 1MQoL 2PQoL 2MQoL 3PQoL 3MQoL
Relationships
1ZCQ2 = 1PQoL
1BVAS = 1PQoL
1PF = 1PQoL
1BP = 1PQoL
1RE = 1PQoL 1MQoL
1MH = 1MQoL
2ZCQ2 = 2PQoL
2BVAS = 2PQoL
2PF = 2PQoL
2BP = 2PQoL
2RE = 2PQoL 2MQoL
2MH = 2MQoL
3ZCQ2 = 3PQoL
3BVAS = 3PQoL
3PF = 3PQoL
3BP = 3PQoL
3RE = 3PQoL 3MQoL
3MH = 3MQoL
Set the Error Covariance of 1PF and 1BVAS Free
Set the Error Covariance of 2PF and 2BVAS Free
Set the Error Covariance of 3PF and 3BVAS Free
Path Diagram
End of Problem

Sample Size =   263

Model 1.2 - respectified Model

        Covariance Matrix

              1ZCQ2      1BVAS       1PF        1BP        1RE        1MH
            --------   --------   --------   --------   --------   --------
   1ZCQ2       0.188
   1BVAS       3.336    670.617
     1PF      -1.582    -44.326     76.373
     1BP      -1.411    -45.800     27.034     67.945
     1RE      -1.760    -67.261     34.215     30.497    204.256
     1MH      -1.383    -28.108     23.101     14.611     99.800    112.878
   2ZCQ2       0.093      1.070     -1.070     -1.204     -1.265     -1.155
   2BVAS       1.682    110.865    -23.107    -39.383    -44.077    -22.354
     2PF      -1.169      1.815     34.938     30.289     24.487     11.868
     2BP      -1.134     -2.959     19.899     33.390     35.690     19.648
     2RE      -0.880    -31.437     12.325     24.570     55.689     33.791
```

2MH	-0.859	-19.296	9.187	11.068	44.273	48.735
3ZCQ2	0.082	2.322	-1.457	-1.249	-0.793	-0.667
3BVAS	2.706	144.768	-54.053	-60.651	-57.101	-42.538
3PF	-1.358	-33.242	40.811	33.234	37.432	25.389
3BP	-0.810	-45.713	23.256	24.363	21.161	17.833
3RE	-0.790	-42.968	19.836	17.306	46.952	27.738
3MH	-0.465	-35.252	8.554	10.866	39.942	43.936

Covariance Matrix

	2ZCQ2	2BVAS	2PF	2BP	2RE	2MH
2ZCQ2	0.409					
2BVAS	8.185	627.355				
2PF	-4.646	-125.244	138.021			
2BP	-5.222	-161.174	90.943	137.439		
2RE	-2.835	-64.153	40.679	50.260	145.449	
2MH	-2.566	-75.977	36.326	52.986	57.323	91.107
3ZCQ2	0.242	8.439	-3.403	-3.714	-2.225	-2.219
3BVAS	7.695	420.235	-135.411	-154.161	-84.674	-95.939
3PF	-3.771	-130.198	86.671	77.477	43.563	45.232
3BP	-3.308	-132.277	62.027	69.313	40.559	41.769
3RE	-2.637	-97.253	34.852	48.137	58.509	46.367
3MH	-1.395	-60.993	19.371	29.559	38.082	48.461

Covariance Matrix

	3ZCQ2	3BVAS	3PF	3BP	3RE	3MH
3ZCQ2	0.440					
3BVAS	13.299	762.301				
3PF	-5.877	-196.328	159.214			
3BP	-6.117	-238.105	109.976	145.638		
3RE	-4.334	-157.400	70.557	75.296	143.668	
3MH	-2.660	-101.739	49.560	48.298	63.244	83.719

Total Variance = 3567.019 Generalized Variance = 0.651428D+28

Largest Eigenvalue = 1511.765 Smallest Eigenvalue = 0.086

Condition Number = 132.363

WARNING: The Condition Number indicates severe multicollinearity.

One or more variables may be redundant.

Model 1.2 - respecified Model

Number of Iterations = 21

LISREL Estimates (Maximum Likelihood)

Measurement Equations

```
   1ZCQ2 = 0.289*1PQoL, Errorvar.= 0.104  , R² = 0.445
Standerr  (0.0284)                (0.0129)
Z-values   10.173                  8.087
P-values    0.000                  0.000

   1BVAS = 9.374*1PQoL, Errorvar.= 577.870, R² = 0.132
Standerr  (1.902)                 (54.951)
Z-values   4.930                   10.516
P-values   0.000                   0.000

     1PF =  - 5.605*1PQoL, Errorvar.= 44.667, R² = 0.413
Standerr     (0.583)                (5.378)
Z-values     -9.621                  8.305
P-values      0.000                  0.000

     1BP =  - 4.918*1PQoL, Errorvar.= 43.497, R² = 0.357
Standerr     (0.545)                (4.757)
Z-values     -9.028                  9.144
P-values      0.000                  0.000

     1RE =  - 2.829*1PQoL + 8.553*1MQoL, Errorvar.= 103.714, R² = 0.490
Standerr     (0.907)       (0.992)                 (12.382)
Z-values     -3.120         8.622                   8.376
P-values      0.002         0.000                   0.000

     1MH = 10.312*1MQoL, Errorvar.= 6.105  , R² = 0.946
Standerr  (0.723)                 (11.256)
Z-values   14.254                  0.542
P-values    0.000                  0.588

   2ZCQ2 = 0.516*2PQoL, Errorvar.= 0.141  , R² = 0.653
```

```
Standarr  (0.0343)                (0.0164)
Z-values   15.039                  8.637
P-values   0.000                   0.000

   2BVAS = 16.509*2PQoL, Errorvar.= 351.172, R² = 0.437
Standarr  (1.490)                 (35.872)
Z-values   11.082                   9.790
P-values   0.000                   0.000

     2PF =  - 9.057*2PQoL, Errorvar.= 55.529, R² = 0.596
Standarr     (0.656)                (6.294)
Z-values     -13.802                 8.823
P-values       0.000                 0.000

     2BP =  - 9.953*2PQoL, Errorvar.= 37.879, R² = 0.723
Standarr     (0.618)                (5.014)
Z-values     -16.101                 7.555
P-values       0.000                 0.000

     2RE =  - 1.709*2PQoL + 5.698*2MQoL, Errorvar.= 98.399, R² = 0.320
Standarr     (0.975)        (1.019)                 (9.728)
Z-values     -1.754          5.592                   10.115
P-values       0.080         0.000                   0.000

     2MH = 8.649*2MQoL, Errorvar.= 16.253, R² = 0.822
Standarr  (0.696)                 (9.122)
Z-values   12.420                  1.782
P-values   0.000                   0.075

   3ZCQ2 = 0.578*3PQoL, Errorvar.= 0.104  , R² = 0.762
Standarr  (0.0329)                (0.0120)
Z-values   17.582                   8.684
P-values   0.000                   0.000

   3BVAS = 22.760*3PQoL, Errorvar.= 241.834, R² = 0.682
Standarr  (1.441)                 (27.454)
Z-values   15.789                   8.809
P-values   0.000                   0.000

     3PF =  - 10.424*3PQoL, Errorvar.= 49.947, R² = 0.685
Standarr     (0.654)                (5.638)
Z-values     -15.946                 8.859
P-values       0.000                 0.000

     3BP =  - 10.398*3PQoL, Errorvar.= 36.964, R² = 0.745
Standarr     (0.602)                (4.128)
Z-values     -17.265                 8.955
P-values       0.000                 0.000

     3RE =  - 4.615*3PQoL + 4.735*3MQoL, Errorvar.= 76.616, R² = 0.465
Standarr     (0.839)        (0.895)                 (7.628)
Z-values     -5.502          5.289                   10.045
P-values       0.000         0.000                   0.000

     3MH = 8.823*3MQoL, Errorvar.= 5.552  , R² = 0.933
Standarr  (0.761)                 (11.308)
Z-values   11.589                  0.491
P-values   0.000                   0.623


Error Covariance for 1PF and 1BVAS = 9.001
                                    (6.130)
                                     1.469

Error Covariance for 2PF and 2BVAS = 23.990
                                     (5.407)
                                      4.437

Error Covariance for 3PF and 3BVAS = 41.589
                                     (4.485)
                                      9.272

        Correlation Matrix of Independant Variables

             1PQoL      1MQoL      2PQoL      2MQoL      3PQoL      3MQoL
          ---------  ---------  ---------  ---------  ---------  ---------
1PQoL       1.000

1MQoL      -0.385      1.000
           (0.071)
           -5.441

2PQoL       0.454     -0.175      1.000
           (0.068)    (0.067)
            6.716     -2.609

2MQoL      -0.271      0.543     -0.560      1.000
           (0.080)    (0.065)    (0.064)
```

	-3.403	8.328	-8.737			
3PQoL	0.468	-0.171	0.724	-0.477	1.000	
	(0.064)	(0.065)	(0.037)	(0.064)		
	7.332	-2.656	19.516	-7.487		
3MQoL	-0.212	0.479	-0.322	0.640	-0.521	1.000
	(0.075)	(0.066)	(0.067)	(0.072)	(0.062)	
	-2.807	7.308	-4.822	8.832	-8.355	

Global Goodness of Fit Statistics, FIML case

-2ln(L) for the saturated model = 29653.620
-2ln(L) for the fitted model = 29935.936

Degrees of Freedom = 114
Full Information ML Chi-Square 282.316 (P = 0.0000)
Root Mean Square Error of Approximation (RMSEA) 0.0749
90 Percent Confidence Interval for RMSEA (0.0640 ; 0.0860)
P-Value for Test of Close Fit (RMSEA < 0.05) 0.000152

The Modification Indices Suggest to Add the

Path to	from	Decrease in Chi-Square	New Estimate
2BVAS	3PQoL	16.3	9.33
3PF	1PQoL	10.4	-2.29
3BP	1PQoL	7.9	1.67

The Modification Indices Suggest to Add an Error Covariance

Between	and	Decrease in Chi-Square	New Estimate
2ZCQ2	1ZCQ2	16.2	0.04
2PF	1PF	14.5	14.45
2BP	1BP	10.0	10.14
2RE	1RE	9.5	20.47
3ZCQ2	1ZCQ2	10.4	0.03
3ZCQ2	2ZCQ2	32.4	0.05
3ZCQ2	2PF	8.2	0.53
3BVAS	2ZCQ2	9.3	-1.38
3BVAS	2BVAS	32.5	120.63
3PF	1PF	10.2	11.60
3PF	2ZCQ2	8.0	0.58
3PF	2PF	28.6	21.47
3RE	1RE	8.8	17.16
3RE	1MH	8.9	-13.11
3RE	2RE	12.3	19.79

Time used 0.593 seconds

Appendix C: Fit Indices for Model 2.1

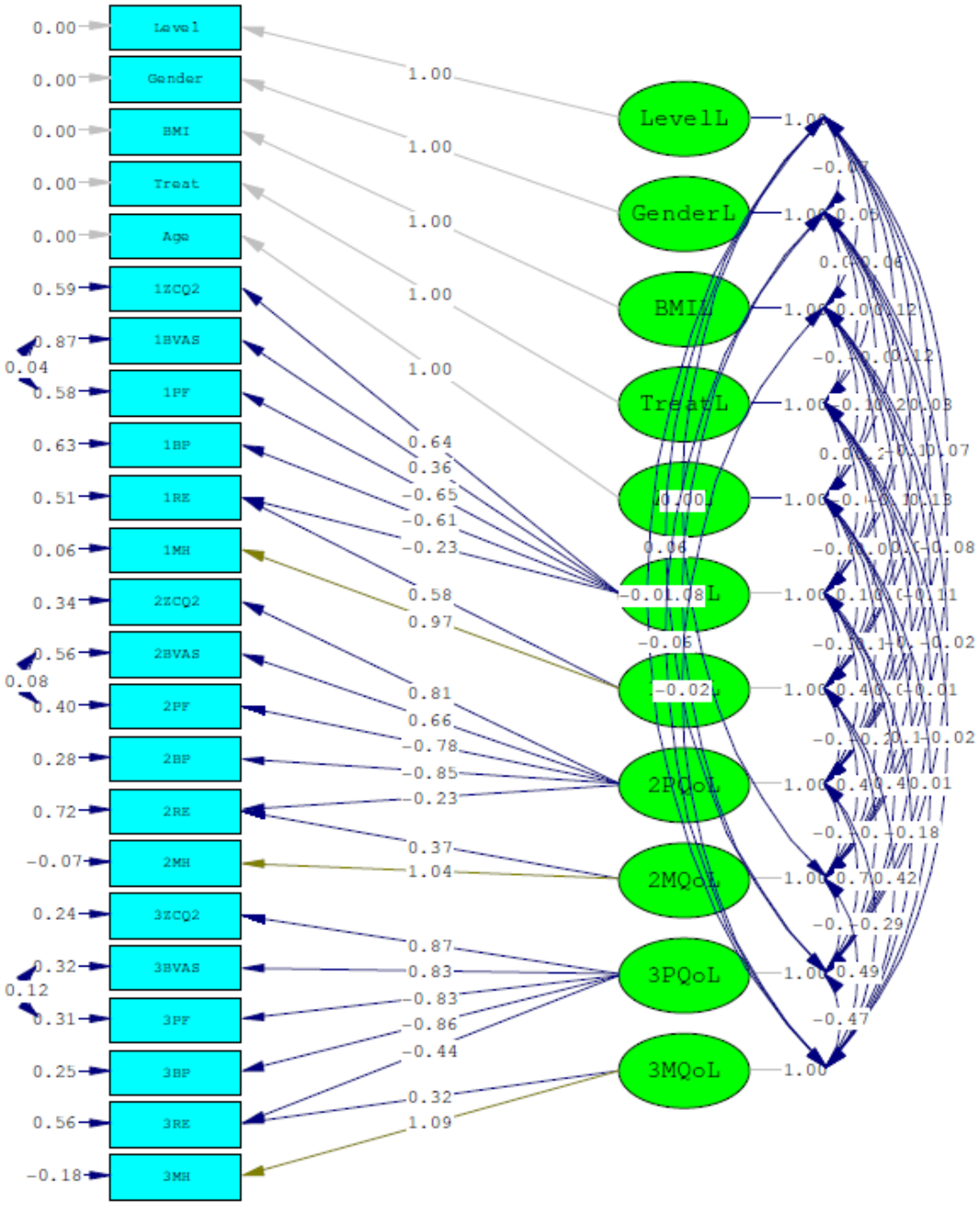
Model	Model Change	χ^2	*df*	RMSEA	χ^2 DIFF	*df* DIFF	*p*
2.1	All variables constrained	426.7	128	0.096			
			Comparison to Model 2.1				
			Level of significance = $\alpha^* = \alpha_f/(n_z n_t)$; 0.05/6 = 0.0083				
2.1.1a	ZCQ2 freed	411.60	126	0.093	15.05	2	0.0005*
2.1.1b	BVAS freed	410.43	126	0.093	16.22	2	0.0003*
2.1.1c	PF freed	417.09	126	0.094	9.56	2	0.0084
2.1.1d	BP freed	405.30	126	0.092	21.35	2	<0.0001*
2.1.1e	RE freed	395.26	124	0.091	31.39	4	<0.0001*
2.1.1f	MH freed	410.96	126	0.093	15.69	2	0.0004*
			Variable RE freed				
			Comparison to Model 2.1.1e				
			Level of significance = $\alpha^* = \alpha_f/(n_z n_t)$; 0.05/5 = 0.01				
2.1.2a	ZCQ2 freed	381.18	122	0.090	14.08	2	0.0009*
2.1.2b	BVAS freed	379.63	122	0.090	15.63	2	0.0004*
2.1.2c	PF freed	386.81	122	0.091	8.43	2	0.0146
2.1.2d	BP freed	375.72	122	0.089	19.54	2	<0.0001*
2.1.2e	MH freed	387.52	122	0.091	7.74	2	0.0209
			Variable BP freed				
			Comparison to Model 2.1.2d				
			Level of significance = $\alpha^* = \alpha_f/(n_z n_t)$; 0.05/4 = 0.0125				
2.1.3a	ZCQ2 freed	350.09	120	0.085	25.63	2	<0.0001*
2.1.3b	BVAS freed	357.34	120	0.087	18.38	2	<0.0001*
2.1.3c	PF freed	361.16	120	0.087	14.56	2	0.0007*
2.1.3d	MH freed	368.36	120	0.089	7.36	2	0.0252
			Variable ZCQ2 freed				
			Comparison to Model 2.1.3a				
			Level of significance = $\alpha^* = \alpha_f/(n_z n_t)$; 0.05/3 = 0.0167				
2.1.4a	BVAS freed	321.94	118	0.081	28.15	2	<0.0001*
2.1.4b	PF freed	321.81	118	0.080	28.28	2	<0.0001*
2.1.4c	MH freed	343.98	118	0.085	6.11	2	0.0471
			Variable PF freed				

Model	Model Change	χ^2	*df*	RMSEA	χ^2 DIFF	*df*	*p*
			Comparison to Model 2.1.4b				
			Level of significance = $\alpha^* = \alpha_f/(n_z n_t)$; 0.05/2 = 0.025				
2.1.5a	BVAS freed	286.05	116	0.075	35.76	2	<0.0001*
2.1.5b	MH freed	316.93	116	0.081	4.88	2	0.0872
			Variable BVAS freed				
			Comparison to Model 2.1.5a				
			Level of significance = $\alpha^* = \alpha_f/(n_z n_t)$; 0.05/1 = 0.05				
2.1.6a	MH freed	282.32	114	0.075	3.73	2	0.1549
			Variable MH not freed - χ^2 DIFF not significant				

Note. ZCQ2 = Zurich Claudication Questionnaire Part 2; BVAS = Back Visual Analog Scale; PF = Physical Function; BP = Bodily Pain; RE = Role Emotional; MH = Mental Health; RMSEA = root mean square error of approximation; χ^2DIFF = chi-square difference; *df* DIFF = degree of freedom difference.

* *p* significant at adjusted level.

Appendix D: Model 3.1F LISREL Standardized Path Diagram and Output File



```
                                DATE:  9/25/2014
                                TIME:  8:03

                          L I S R E L  9.10 (32 Bit)

                                     BY

                          Karl G. Jöreskog & Dag Sörbom


                       This program is published exclusively by
                       Scientific Software International, Inc.
                            http://www.ssicentral.com

             Copyright by Scientific Software International, Inc., 1981-2012
               Use of this program is subject to the terms specified in the
                                 Universal Copyright Convention.

The following lines were read from file C:\Users\rsc\Desktop\Research\Model 3.1.spj:

Model 3.1 - Exogenous Variables
Raw Data from file 'C:\Users\rsc\Desktop\Research\RSData_Longitudinal.LSF'

                 ------------------------------
                  EM Algorithm for missing Data:
                 ------------------------------

               Number of different missing-value patterns=        5
               Effective sample size:      263

               Convergence of EM-algorithm in     4 iterations
               -2 Ln(L) =    34112.72533
               Percentage missing values=   1.44

          Note:
            The Covariances and/or Means to be analyzed are estimated
            by the EM procedure and are only used to obtain starting
            values for the FIML procedure

Sample Size = 263
Latent Variables  LevelL GenderL BMIL TreatL AgeL 1PQoL 1MQoL 2PQoL 2MQoL 3PQoL 3MQoL
Relationships
Level = 1.0*LevelL
Gender = 1.0*GenderL
BMI = 1.0*BMIL
Treat = 1.0*TreatL
Age = 1.0*AgeL
1ZCQ2 = 1PQoL
1BVAS = 1PQoL
1PF = 1PQoL
1BP = 1PQoL
1RE = 1PQoL 1MQoL
1MH = 1MQoL
2ZCQ2 = 2PQoL
2BVAS = 2PQoL
2PF = 2PQoL
2BP = 2PQoL
2RE = 2PQoL 2MQoL
2MH = 2MQoL
3ZCQ2 = 3PQoL
3BVAS = 3PQoL
3PF = 3PQoL
3BP = 3PQoL
3RE = 3PQoL 3MQoL
3MH = 3MQoL
Set the Error Variance of Level to 0.00
Set the Error Variance of Gender to 0.00
Set the Error Variance of BMI to 0.00
Set the Error Variance of Treat to 0.00
Set the Error Variance of Age to 0.00
Set the Error Covariance of 1PF and 1BVAS Free
Set the Error Covariance of 2PF and 2BVAS Free
Set the Error Covariance of 3PF and 3BVAS Free
Set the Path from 2MQoL to 2MH and the Path from 1MQoL to 1MH Equal
Set the Path from 3MQoL to 3MH and the Path from 1MQoL to 1MH Equal
Set the Path from 3MQoL to 3MH and the Path from 2MQoL to 2MH Equal
Path Diagram
End of Problem

Sample Size =   263

Model 3.1 - Exogenous Variables

        Covariance Matrix
```

Level	0.250					
Gender	-0.017	0.235				
BMI	0.118	0.042	22.723			
Age	0.561	-0.165	-6.288	0.224	92.079	
1ZCQ2	0.018	0.020	0.282	-0.021	-0.010	0.188
1BVAS	1.178	1.574	4.815	0.254	-4.560	3.329
1PF	-0.339	-0.637	-8.135	0.050	-1.526	-1.582
1BP	-0.136	-0.854	-4.710	0.407	4.774	-1.411
1RE	-0.373	-1.906	-7.015	-0.040	5.571	-1.760
1MH	0.155	-0.683	-4.866	0.310	18.859	-1.383
2ZCQ2	0.013	0.035	0.130	0.018	1.052	0.093
2BVAS	0.806	1.015	-5.283	0.611	21.817	1.695
2PF	-0.438	-0.740	-5.702	-0.254	-19.703	-1.174
2BP	-0.331	-0.582	-4.490	-0.064	-12.435	-1.142
2RE	-0.212	0.158	-5.039	-0.260	-16.071	-0.890
2MH	-0.003	-0.382	-5.376	-0.135	7.254	-0.865
3ZCQ2	0.028	0.017	-0.107	-0.005	1.128	0.082
3BVAS	0.477	1.036	-9.633	-0.024	21.018	2.728
3PF	-0.336	-0.552	-3.205	0.279	-26.308	-1.358
3BP	-0.037	-0.305	1.342	-0.091	-17.231	-0.810
3RE	-0.378	-0.059	-4.400	0.053	-18.448	-0.790
3MH	-0.083	-0.295	-1.199	-0.089	0.706	-0.465

Covariance Matrix

	1BVAS	1PF	1BP	1RE	1MH	2ZCQ2
1BVAS	669.899					
1PF	-44.132	76.373				
1BP	-45.704	27.034	67.945			
1RE	-66.776	34.215	30.497	204.256		
1MH	-27.721	23.101	14.611	99.800	112.878	
2ZCQ2	1.083	-1.074	-1.227	-1.291	-1.153	0.410
2BVAS	111.131	-23.472	-39.693	-44.585	-22.451	8.240
2PF	1.534	35.003	30.829	25.005	11.785	-4.672
2BP	-2.640	19.874	33.777	35.975	19.588	-5.248
2RE	-30.440	12.395	24.960	56.037	33.828	-2.857
2MH	-18.919	9.187	11.215	44.381	48.759	-2.582
3ZCQ2	2.312	-1.457	-1.249	-0.793	-0.667	0.243
3BVAS	144.809	-54.353	-60.763	-57.485	-42.724	7.755
3PF	-33.134	40.811	33.234	37.432	25.389	-3.771
3BP	-45.420	23.256	24.363	21.161	17.833	-3.306
3RE	-42.592	19.836	17.306	46.952	27.738	-2.650
3MH	-34.993	8.554	10.866	39.942	43.936	-1.391

Covariance Matrix

	2BVAS	2PF	2BP	2RE	2MH	3ZCQ2
2BVAS	629.096					
2PF	-126.306	138.418				
2BP	-161.837	91.417	137.881			
2RE	-64.671	41.360	50.752	145.601		
2MH	-76.152	36.407	53.157	57.491	91.318	
3ZCQ2	8.458	-3.423	-3.728	-2.223	-2.236	0.440
3BVAS	421.644	-136.484	-154.679	-85.475	-95.962	13.306
3PF	-130.519	86.623	77.467	43.918	45.162	-5.877
3BP	-132.392	62.009	69.273	40.435	41.973	-6.117
3RE	-97.906	35.186	48.283	58.573	46.656	-4.334
3MH	-61.143	19.300	29.416	37.696	48.573	-2.660

Covariance Matrix

	3BVAS	3PF	3BP	3RE	3MH
3BVAS	762.927				
3PF	-196.434	159.214			
3BP	-238.237	109.976	145.638		
3RE	-157.682	70.557	75.296	143.668	
3MH	-101.945	49.560	48.298	63.244	83.719

Total Variance = 3685.405 Generalized Variance = 0.104665D+30

Largest Eigenvalue = 1516.181 Smallest Eigenvalue = 0.085

Condition Number = 133.830

WARNING: The Condition Number indicates severe multicollinearity.

One or more variables may be redundant.

Model 3.1 - Exogenous Variables

```
   Level = 1.000*LevelL, , R² = 1.000

  Gender = 1.000*GenderL, , R² = 1.000

     BMI = 1.000*BMIL, , R² = 1.000

   Treat = 1.000*TreatL, , R² = 1.000

     Age = 1.000*AgeL, , R² = 1.000

   1ZCQ2 = 0.276*1PQoL, Errorvar.= 0.111  , R² = 0.407
Standerr  (0.0279)                (0.0126)
Z-values   9.875                   8.802
P-values   0.000                   0.000

   1BVAS = 9.266*1PQoL, Errorvar.= 578.930, R² = 0.129
Standerr  (1.870)                 (54.574)
Z-values   4.956                   10.608
P-values   0.000                   0.000

     1PF =  - 5.639*1PQoL, Errorvar.= 44.074, R² = 0.419
Standerr     (0.569)                  (5.194)
Z-values     -9.905                    8.485
P-values      0.000                    0.000

     1BP =  - 4.993*1PQoL, Errorvar.= 42.591, R² = 0.369
Standerr     (0.535)                  (4.627)
Z-values     -9.341                    9.206
P-values      0.000                    0.000

     1RE =  - 3.198*1PQoL + 8.177*1MQoL, Errorvar.= 102.297, R² = 
Standerr     (0.866)       (0.952)                  (11.382)
Z-values     -3.691         8.587                    8.987
P-values      0.000         0.000                    0.000

     1MH = 10.024*1MQoL, Errorvar.= 5.989 , R² = 0.944
Standerr  (0.542)                 (9.149)
Z-values   18.498                   0.655
P-values   0.000                   0.513

   2ZCQ2 = 0.521*2PQoL, Errorvar.= 0.140  , R² = 0.659
Standerr  (0.0343)                (0.0163)
Z-values   15.189                   8.630
P-values   0.000                   0.000

   2BVAS = 16.566*2PQoL, Errorvar.= 352.919, R² = 0.437
Standerr  (1.490)                 (35.948)
Z-values   11.118                   9.818
P-values   0.000                   0.000

     2PF =  - 9.145*2PQoL, Errorvar.= 55.011, R² = 0.603
Standerr     (0.655)                  (6.239)
Z-values     -13.964                   8.817
P-values      0.000                    0.000

     2BP =  - 9.983*2PQoL, Errorvar.= 38.554, R² = 0.721
Standerr     (0.619)                  (5.019)
Z-values     -16.123                   7.682
P-values      0.000                    0.000

     2RE =  - 2.772*2PQoL + 4.514*2MQoL, Errorvar.= 105.396, R² = 
Standerr     (0.866)       (0.913)                  (9.733)
Z-values     -3.199         4.945                   10.828
P-values      0.001         0.000                    0.000

     2MH = 10.024*2MQoL, Errorvar.= -6.718  , R² = 1.072
Standerr  (0.542)                 (10.999)
Z-values   18.498                  -0.611
P-values   0.000                   0.541

W_A_R_N_I_N_G : Error variance is negative.

   3ZCQ2 = 0.581*3PQoL, Errorvar.= 0.105  , R² = 0.763
Standerr  (0.0330)                (0.0120)
Z-values   17.631                   8.740
P-values   0.000                   0.000

   3BVAS = 22.895*3PQoL, Errorvar.= 241.707, R² = 0.684
Standerr  (1.442)                 (27.408)
Z-values   15.872                   8.819
P-values   0.000                   0.000
```

```
Standerr      (0.653)                  (5.588)
Z-values      -16.089                   8.838


     3BP =  - 10.444*3PQoL, Errorvar.= 37.201, R² = 0.746
Standerr      (0.603)                  (4.131)
Z-values      -17.313                   9.006
P-values       0.000                    0.000

     3RE =  - 5.329*3PQoL + 3.890*3MQoL, Errorvar.= 81.450, R² = 0.436
Standerr      (0.725)        (0.746)                 (7.331)
Z-values      -7.349          5.218                   11.110
P-values       0.000          0.000                   0.000

     3MH = 10.024*3MQoL, Errorvar.= -15.324 , R² = 1.180
Standerr   (0.542)                  (11.209)
Z-values    18.498                  -1.367
P-values    0.000                    0.172

W_A_R_N_I_N_G : Error variance is negative.


Error Covariance for 1PF and 1BVAS = 9.272
                                    (5.994)
                                     1.547

Error Covariance for 2PF and 2BVAS = 24.039
                                    (5.387)
                                     4.462

Error Covariance for 3PF and 3BVAS = 42.222
                                    (4.462)
                                     9.462
```

Covariance Matrix of Independent Variables

	LevelL	GenderL	BMIL	TreatL	AgeL	1PQ
LevelL	0.249 (0.022) 11.467					
GenderL	-0.017 (0.015) -1.154	0.234 (0.020) 11.468				
BMIL	0.118 (0.146) 0.802	0.041 (0.142) 0.287	22.633 (1.974) 11.467			
TreatL	-0.016 (0.015) -1.043	0.017 (0.015) 1.121	-0.081 (0.146) -0.553	0.248 (0.022) 11.467		
AgeL	0.557 (0.296) 1.881	-0.158 (0.285) -0.553	-6.227 (2.833) -2.198	0.219 (0.294) 0.746	91.555 (7.983) 11.468	
1PQoL	0.062 (0.037) 1.689	0.132 (0.035) 3.758	1.091 (0.347) 3.147	-0.039 (0.037) -1.067	-0.034 (0.708) -0.049	1.0
1MQoL	0.013 (0.032) 0.413	-0.068 (0.031) -2.232	-0.460 (0.302) -1.524	0.027 (0.032) 0.845	1.724 (0.607) 2.839	-0.3 (0.07 -5.2
2PQoL	0.034 (0.033) 1.015	0.064 (0.032) 1.996	0.325 (0.315) 1.033	0.026 (0.033) 0.788	1.714 (0.629) 2.725	0.4 (0.06 6.8
2MQoL	0.002 (0.030) 0.050	-0.041 (0.029) -1.388	-0.530 (0.289) -1.836	-0.015 (0.030) -0.494	0.732 (0.578) 1.266	-0.2 (0.07 -3.0
3PQoL	0.028 (0.032) 0.868	0.039 (0.031) 1.264	-0.075 (0.304) -0.247	-0.007 (0.032) -0.224	1.806 (0.606) 2.983	0.4 (0.06 7.4
3MQoL	-0.007 (0.028) -0.259	-0.031 (0.027) -1.127	-0.085 (0.269) -0.317	-0.010 (0.028) -0.370	0.106 (0.541) 0.197	-0.1 (0.06 -2.7

Covariance Matrix of Independent Variables

	1MQoL	2PQoL	2MQoL	3PQoL	3MQoL

2PQoL	-0.180	1.000			
2MQoL	0.470	-0.498	1.000		
	(0.064)	(0.059)			
	7.353	-8.387			
3PQoL	-0.179	0.727	-0.428	1.000	
	(0.064)	(0.037)	(0.058)		
	-2.796	19.890	-7.349		
3MQoL	0.423	-0.295	0.494	-0.471	1.000
	(0.061)	(0.058)	(0.068)	(0.055)	
	6.944	-5.069	7.282	-8.551	

Global Goodness of Fit Statistics, FIML case

-2ln(L) for the saturated model = 34112.725
-2ln(L) for the fitted model = 34483.048

Degrees of Freedom = 176
Full Information ML Chi-Square 370.322 (P = 0.
Root Mean Square Error of Approximation (RMSEA) 0.0648
90 Percent Confidence Interval for RMSEA (0.0556 ; 0.074
P-Value for Test of Close Fit (RMSEA < 0.05) 0.00488

The Modification Indices Suggest to Add the

Path to	from	Decrease in Chi-Square	New Estimate
1RE	GenderL	10.6	-4.63
1MH	GenderL	8.9	5.18
2BVAS	3PQoL	16.2	9.39
2MH	AgeL	19.1	1.16
2MH	1PQoL	16.4	14.83
3PF	1PQoL	11.2	-2.34

The Modification Indices Suggest to Add an Error Covariance

Between	and	Decrease in Chi-Square	New Estimate
1RE	Gender	10.5	-0.99
1MH	Gender	10.4	1.21
2ZCQ2	1ZCQ2	16.0	0.04
2PF	1PF	12.4	13.20
2BP	1BP	10.5	10.31
2RE	1RE	10.0	20.85
2MH	Gender	9.7	-3.40
2MH	Age	23.9	106.75
3ZCQ2	1ZCQ2	9.3	0.03
3ZCQ2	2ZCQ2	32.1	0.05
3ZCQ2	2PF	8.0	0.52
3BVAS	2ZCQ2	8.8	-1.33
3BVAS	2BVAS	33.7	122.75
3PF	1PF	9.6	11.11
3PF	2ZCQ2	8.9	0.61
3PF	2PF	28.4	21.18
3RE	1RE	8.8	17.14
3RE	2RE	11.1	18.78

Time used 1.591 seconds

Time used 1.591 seconds

Curriculum Vitae

Robin Sneed Carlson, Ph.D.

EDUCATION:

Doctor of Philosophy - Applied Management and Decision Sciences, Operations Research, Walden University, Minneapolis, MN — 2015

Master of Business Administration
University of Phoenix, Phoenix, AZ — 1995

Bachelor of Science - Electrical Engineering
U.S. Air Force Academy, Colorado Springs, CO — 1983

Relevant Professional Experience:

Sr. Manager, Clinical Affairs — 2012 –2014
Ventana Medical Systems**,** Tucson, AZ
Managed companion diagnostic clinical studies for in vitro assays including coordination with development, regulatory, medical office, manufacturing, and biostatistics. Initiated and coordinated studies with laboratories both in the United States and internationally.

Director, Clinical Research — 2010 –2011
VertiFlex**,** San Clemente, CA
Management of start-up company's primary IDE clinical trial of an interspinous spacer with emphasis on CRA and database management. Supervise four clinical staff and responsible for four support vendor contracts. Provide clinical strategic planning, board support, predictive trending, and resolve site study issues.

Director of Engineering — 2009
U.S. Air Force, Hill AFB, UT
Provided strategic vision, mentoring, and oversight to 50 engineers as a voluntarily activated Reserve Officer. Senior engineering leader for the sustainment of 22 U.S. Air Force programs, ensured standard procedures were consistently and effectively applied to all processes. Led the successful merger of two fundamentally different organizations during restructuring by analyzing issues, devising solutions, and communicating results to all audiences.

Director, Clinical Affairs — 2004 –2008
EKOS Corporation, Bothell, WA
As part of the Senior Management Team of a start-up company provided clinical expertise/strategy and designed and managed multiple post marketing device studies. Supervised two CRAs and responsible for budget of $650,000. Led marketing tasks including video production, journal article/ abstract preparation, and conference support. Supported production with accurate forecasts and engineering with requirement definitions.

Clinical Project Manager 2000 –2004
W.L. Gore & Associates, Flagstaff, AZ
Designed, initiated, conducted and documented clinical trials on implantable devices including thoracic aneurysm stent-grafts and vascular bypass grafts. Led team of 14 and budget of over $1 Million. Provided data and analysis to support FDA submissions and peer-reviewed journal articles. Succeeded in unique corporate environment with no titles or supervisors.

Project Manager, Medical Devices 1997 –2000
Quintiles, Inc., Rockville, MD
Project Manager for a Contract Research Organization specializing in medical device clinical trials. Managed multiple projects for a variety of clients, supervised 4 clinical research associates and 7 monitors and coordinated with internal departments (data management, regulatory, etc.) to product results.

Sr. Clinical Programs Associate 1996 –1997
Mentor Corporation, Santa Barbara, CA
Supported clinical studies in plastic surgery, ophthalmics and urology by monitoring clinical sites, performing data quality reviews, analyzing data and writing comprehensive final/annual reports.

Clinical Program Manager 1990 –1996
Intermedics, Inc., Angleton, TX
Designed and managed clinical evaluations on implantable cardiac products for FDA submissions and marketing requirements. Provided technical support to physicians, pacemaker clinics, patients and sales representatives.

Product Specialist 1989 –1990
Siecor Corporation, Keller, TX
Provided technical information to customers and the sales force on fiber optic network interface products. Assisted in the improvement of manufacturing operations and development of new products.

Communications Officer 1983 –1989
United States Air Force, Wheeler AFB, HI & Wright-Patterson AFB, OH
Engineered installation of communications equipment in the Pacific Theater and led operational test teams. Developed secure voice and satellite systems test plans, supervised 10 test engineers and led evaluations.

CERTIFICATION

Regulatory Affairs Certified (RAC) 1995-present

PUBLICATIONS:

Sneed, R. & Kilmer, R. (2012). The Air Force's Individual Mobilization Augmentee Program: Is the Current Organizational Structure Viable? *Air & Space Power Journal, 26*(5), 12-32. Retrieved from http://www.airpower.maxwell.af.mil

Carlson, R. (2007). *EKOS Ultrasound Accelerated Thrombolysis: Venous Occlusions – Deep Vein Thrombolysis* (White Paper). Bothell, WA: EKOS Corporation.

Carlson, R. (2007). *EKOS Ultrasound Accelerated Thrombolysis: Peripheral Arterial Occlusions – Early Experience* (White Paper). Bothell, WA: EKOS Corporation.

Rizo-Patron,C., Belott,P., Belco,K., Zhu,W., Zinner,A. & Sneed, R. (1995). *Initial Multicenter Experience with New Thin Bipolar Pacing Lead* (Abstract). *Pacing and Clinical Electrophysiology, 18*(4, Part II).

Singer, I., Peavler, P., Johnson, B. & Sneed, R. (1993). Temperature May Be An Ideal Sensor for Rate Modulation in Patients with Posture-Related Syncope and Chronotropic Incompetence (Abstract). *Proceedings of the American College of Cardiology, 42nd Scientific Session*, USA.

PROFESSIONAL AFFILIATIONS

Member, Regulatory Affairs Professional Society
Member, INFORMS