12-1-2018

Shoulder-Specific Patient Reported Outcome Measures for Use in Patients with Head and Neck Cancer: An Assessment of Reliability, Construct Validity, and Overall Appropriateness of Test Score Interpretation Using Rasch Analysis

Melissa Michelle Eden
Nova Southeastern University

Shoulder-Specific Patient Reported Outcome Measures for Use in Patients with Head and Neck Cancer: An Assessment of Reliability, Construct Validity, and Overall Appropriateness of Test Score Interpretation Using Rasch Analysis

by

Melissa M. Eden PT, DPT, OCS

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Nova Southeastern University
Dr. Pallavi Patel College of Health Care Sciences
Department of Physical Therapy
2018

Approval/Signature Page

We hereby certify that this dissertation, submitted by Melissa M. Eden, conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirement for the degree of Doctor of Philosophy in Physical Therapy.

_____ _____

Dr. M. Samuel Cheng, PT, MS, ScD — Date
Chairperson of Dissertation Committee

_____ _____

Dr. Morey J. Kolber, PT, PhD, OCS — Date
Member, Dissertation Committee

_____ _____

Dr. Mary Lou Galantino, PT, MS, PhD, MSCE, FAPTA — Date
Member, Dissertation Committee

Approved:

_____ _____

Dr. M. Samuel Cheng, PT, MS, ScD — Date
Director, Physical Therapy Ph.D Program

_____ _____

Dr. Shari Rone-Adams PT, MHSA, DBA — Date
Chair, Department of Physical Therapy

_____ _____

Dr. Stanley H. Wilson, PT, EdD, CEAS — Date
Dean and Associate Professor

Abstract

Context: Medical management for head and neck cancer (HNC) often includes neck dissection surgery, a side effect of which is shoulder dysfunction. There is no consensus for which patient-reported outcome measure (PRO) is most appropriate to quantify shoulder dysfunction in this population.

Objective: The aims of this research study were to: (1) use Rasch methodologies to assess construct validity and overall appropriateness of test score interpretation of Disability of the Arm, Shoulder and Hand (DASH), QuickDASH, Shoulder Pain and Disability Index (SPADI) and Neck Dissection Impairment Index (NDII) in the HNC population; (2) determine appropriateness of use of University of Washington Quality of Life (UW-QoL) shoulder subscale as a screening tool for shoulder impairment; (3) recommend a new PRO, or combination of PROs, that more accurately portrays the construct of shoulder dysfunction in the HNC population.

Design: One hundred and eight-two individuals who had received a neck dissection procedure within the past 2 weeks to 18 months completed the PROs. Rasch methodologies were utilized to address the primary aim of the study through consideration of scale dimensionality [principal components analysis, item and person fit, differential item functioning (DIF)], scale hierarchy (gaps/redundancies, floor/ceiling effects, coverage of ability levels), response scale structure, and reliability (person and item reliability and separation statistics). The secondary aim was addressed through correlational analysis of the UW-QoL (shoulder subscale), DASH, QuickDASH, SPADI and NDII.

Results: The DASH did not meet criteria for unidimensionality, and was deemed inappropriate for utilization in this sample. The QuickDASH, SPADI and NDII were all determined to be

unidimensional. All scales had varying issues with person and item misfit, DIF, coverage of ability levels, gaps/redundancies, and optimal rating scale requirements. The NDII meets most requirements. All measures were found to meet thresholds for person and item separation and reliability statistics. The third aim of this study was not addressed because the NDII was determined to be appropriate for this population.

Conclusions: Rasch analysis indicates the NDII is the most appropriate measure studied for this population. The QuickDASH and SPADI are recommended with reservation. The DASH and the UW-QoL (shoulder subscale) are not recommended.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Introduction to the Chapter

This chapter will provide an introduction to the study, including a statement of the problem that was investigated and aims of the investigation. This chapter also provides a rationale for the relevance and need for the study, a list of study questions, and definitions of key terms.

Problem Statement and Study Aims

A patient with a diagnosis of head and neck cancer (HNC) often requires a neck dissection procedure to remove the lymph nodes from the neck, a common site of metastasis. This procedure may damage the spinal accessory nerve (SAN) resulting in trapezius muscle weakness or atrophy and impaired shoulder mobility.[1] Although there are nearly fifty patient-reported outcome measures (PROs) related to shoulder function used in the literature,[2] there is no widely accepted PRO for patients presenting with shoulder dysfunction following neck dissection surgery for HNC.

In 2013, Goldstein and colleagues published a review of the six PROs used in the research literature to quantify shoulder function in patients with HNC. Only one of the measures, the Neck Dissection Impairment Index (NDII), was specifically designed for use in HNC.[3] The NDII, a measure of quality of life (QOL), must be used with caution because it was not developed using sound methodology and has limited testing of its psychometric properties.[3,4] The remaining five PROs described in the review, the Shoulder Pain and Disability Index (SPADI), Shoulder Disability Questionnaire (SDQ), Constant's Shoulder Score (CS), American Shoulder and Elbow Surgeons Standardized Form (ASES), and the Disabilities of the Arm,

Shoulder and Hand Questionnaire (DASH), must also be used with caution. Although these measures have demonstrated strong validity and responsiveness across musculoskeletal populations, they have not been adequately tested in patients with HNC.[3]

The Academy of Oncologic Physical Therapy of the American Physical Therapy Association (APTA) established a Task Force as part of the Evaluation Database to Guide Effectiveness (EDGE) initiative to explore the appropriateness of various outcome measures for use in HNC. A systematic literature review specific to shoulder-related PROs has been completed.[2] Using pre-established criteria for recommendations, the Task Force recommends the DASH and its shortened version the QuickDASH, the NDII, SPADI, and a shoulder subscale of the University of Washington Quality of Life scale (UW-QoL). Although the Task Force took steps to decrease potential bias, the validity of the recommendations must be questioned because of limitations in methodology, limited research related to treatment intervention, minimal research establishing the appropriateness of PROs specific to the HNC population, and the Task Force's reliance on reporting psychometrics based on the traditional Classical Test Theory (CTT). Outcome measures analyzed using CTT allow for valid-population based research, however must be generalized to the individual patient with caution.[5]

Research Study Aims

The aims of this study were three-fold:

- Use Rasch methodology to assess the construct validity and overall appropriateness of test score interpretation of the DASH, QuickDASH, SPADI, and the NDII in patients experiencing shoulder dysfunction following neck dissection surgery for HNC; and
- Determine the appropriateness of use of the UW-QoL (shoulder subscale) as a screening tool for shoulder-related impairment.

- Based on these findings, suggest a combination of outcome measures, or a new outcome measure, that more accurately portrays shoulder disability in patients who experience shoulder dysfunction following a neck dissection surgery for HNC

Relevance, Significance, and Need for the Study

Rasch analysis is gaining popularity in physical therapy research because it allows for a more thorough psychometric analysis of an instrument on a test item or individual level.[5] Recent studies have analyzed frequently used PROs, including the DASH and QuickDASH, using Rasch analysis resulting in a better understanding of how the outcome measure functions on an item level, and its construct validity.[6-18]

Rasch analysis of the DASH and QuickDASH suggests limitations in construct validity related to item fit, dimensionality, item response option thresholds, response scale structure, and residual correlations in patients with musculoskeletal complaints, patients with multiple sclerosis, stroke, and in patients with Dupuytren's contracture.[10,11,13,14,18] The usefulness of the QuickDASH has been further questioned based on unresolved weaknesses found in the analysis of the DASH specific to misfit of two items, tingling and sexual activity, one of which, tingling, is also used in the QuickDASH.[10,14] Preliminary Rasch analysis of the DASH and QuickDASH in 131 patients with HNC confirms limitations of the ability of both scales to measure mild to moderate disability levels in patients presenting with shoulder dysfunction following unilateral neck dissection. In addition, both tools demonstrate problems with item misfit, test item redundancy, and ceiling effects.[19]

The SPADI was analyzed using Rasch analysis in a population of surgical and non-surgical patients presenting to a private practice orthopedic surgery clinic with upper extremity complaints. The authors report misfit of three test items. In addition, the SPADI was found to

have poor precision measuring shoulder dysfunction at the low and high ends of the scale. The tool, however, demonstrates good precision in patients with mid-range function indicating that the items tend to measure mid-level ability related to shoulder dysfunction.[20]

The Task Force chose to recommend the UW-QoL (shoulder subscale) and the NDII because they were developed specifically for the HNC population with shoulder dysfunction and have shown promising reliability and validity using CTT methodology. Stuiver and colleagues used a one-parameter logistic Rasch model to study the psychometric properties of a measure that combines the SPADI and the NDII in a population of 107 subjects who were within one to eight months from neck dissection surgery. Rasch analysis supported the unidimensionality of the combined scales, but showed disordered response scale structure, gaps in scale hierarchy, and redundancies.[21] The psychometric properties of the SPADI, and NDII have not been individually assessed using Rasch analysis.

The use of Rasch analysis to study the psychometric properties and scale functioning of the DASH, QuickDASH, SPADI, and the NDII will provide the physical therapist with a better understanding of how the tools function when used in the HNC population, in addition to their construct validity, and the appropriateness of test score interpretation. In addition, Rasch analysis will provide a better understanding of which PRO, or combination of PROs, are best suited to quantify shoulder function in patients undergoing neck dissection for HNC. The single-item nature of the UW-QoL (shoulder subscale) prohibits its inclusion in the Rasch analysis, however further assessment of his relationship and usability as a screening tool for physical therapy would be of benefit.

The current culture of health care requires a healthcare provider to demonstrate value, which is determined by the benefit of the intervention (change) divided by the cost to provide

that intervention. A healthcare provider must accurately measure change in a patient's condition resulting from the intervention provided. Although PROs have been in existence for decades, recent Medicare mandates for functional outcome reporting have increased the awareness and use of PROs within the physical therapy profession.[22] Physical therapists can use PROs to quantify change resulting from an intervention. The appropriate use and interpretation of PROs requires that adequate construct validity, reliability, and responsiveness to change have been demonstrated in the population of interest. Many of the PROs available for use in the HNC population have not been appropriately researched to allow for accurate use and interpretation of function and change scores.

The APTA's EDGE Task Force on Head and Neck Cancer has recommended the DASH, QuickDASH, SPADI, NDII, and the shoulder subscale of the UW-QoL for use in this patient population.[2] This research study will provide the physical therapist with an unbiased, statistically sound understanding of the extent that these PRO test scores can be utilized in quantifying disability related to shoulder dysfunction in patients following a neck dissection procedure for HNC. If the analysis shows that the measures are not appropriate, recommendations will be made as to which combination of measures, or whether a newly developed measure, would be more acceptable.

Research Questions to be Investigated

To answer the primary research question, "Which of the recommended outcome measures demonstrates acceptable psychometric characteristics allowing for accurate test score interpretation in patients presenting with shoulder disability in the setting of HNC?," construct validity and reliability of the DASH, QuickDASH, SPADI and the NDII will be analyzed using Rasch methodologies for the following investigational questions:

Questions for Construct Validity

Scale Dimensionality

1. Does Principle Component Analysis (PCA) support the assumption of unidimensionality?
2. Are there sub-scales within the measure that should be considered for analysis separately?
3. Is there test item misfit to suggest the presence of additional constructs within the PRO?
4. Is there person misfit to suggest flaws in the intended item hierarchy?
5. Do individuals answer test items differently based on age or gender (DIF)?

Scale Hierarchy

1. Does the scale hierarchy cover the entire spectrum of the construct of shoulder dysfunction?
2. Are there gaps and/or redundancies?
3. Does the scale hierarchy demonstrate the presence of floor and/or ceiling effects?
4. Which test items are considered to be the easiest or most difficult?

Response Scale Structure

1. Are there at least 10 responses per response option category?
2. Are response options equally utilized?
3. Are there disordered response options or step calibrations?
4. Are the average measures ordered?
5. Are there response category outfit MNSQ values that exceed 2.0?
6. Does collapsing the response categories improve:
    a. Item and person reliability and separation indices,

b. Response option utilization,

c. Ordering (step calibration, average measures), and

d. Outfit MNSQ values?

Reliability

1. Does person separation suggest a good ability to separate individuals based upon ability level?
2. Does item separation verity item hierarchy?
3. Does person reliability meet expected cut-off values for individual-level analysis?
4. Does item reliability verify item hierarchy?

CTT methodologies, including correlational analysis of the UW-QoL (shoulder subscale), will be utilized to address the usability of the single test item as a screening tool for shoulder dysfunction in this population.

Guide to Abbreviations for the Dissertation

ADL	Activities of Daily Living
APTA	American Physical Therapy Association
ASES	Modified American Shoulder and Elbow Surgeons standardized form
AUC	Area under the curve
BMI	Body mass index
cASES	Modified American Shoulder and Elbow Surgeons standardized form – clinician completed section
CAT	Computerized Adaptive Testing
CI	Confidence interval
cm	Centimeter
CMS	Centers for Medicare and Medicaid Services
CS	Constant's Shoulder Score
CT	Computed tomography
CTT	Classical Test Theory
DASH	Disabilities of the Arm, Shoulder and Hand
DIF	Differential item functioning
EDGE	Evaluation Database to Guide Effectiveness
EHR	Electronic health record

EMG	Electromyography
EORTC	European Organisation for Research and Treatment of Cancer
ES	Effect size
HIPAA	Health Insurance Portability and Accountability Act
HNC	Head and neck cancer
HPV	Human papilloma virus
HRQOL	Health-related quality of life
ICC	Intraclass correlation coefficient
ICF	International Classification of Human Functioning and Health
IJV	Internal jugular vein
IRB	Institutional Review Board
IRT	Item Response Theory
MCID	Minimal clinically important difference
MDC	Minimal detectable change
MID	Minimal important difference
MNSQ	Mean-square
MRND	Modified radical neck dissection
MS	Multiple Sclerosis
NDII	Neck Dissection Impairment Index
OA	Osteoarthritis
pASES	Modified American Shoulder and Elbow Surgeons standardized form – patient reported section
PCA	Principle component analysis
PRET	Progressive resistance exercise training
PRO	Patient-reported outcome measure
PROMIS®	Patient Reported Outcomes Measurement Information System
QOL	Quality of life
RA	Rheumatoid arthritis
RCT	Randomized controlled trial
RND	Radical neck dissection
ROC-curve	Receiver operating characteristic curve
ROM	Range of motion
r	Spearman's and Person's correlation coefficient
SAN	Spinal accessory nerve
SCCA	Squamous cell carcinoma
SCM	Sternocleidomastoid muscle
SD	Standard deviation
SDQ	Shoulder Disability Questionnaire
SDQ-NL	Shoulder Disability Questionnaire-Netherlands
SDQ-UK	Shoulder Disability Questionnaire-United Kingdom
SE	Standard error
SEM	Standard error of the measure
SF-36	Medical Outcomes Short Form 36
SIP	Sickness Impact Profile
SND	Selective neck dissection

SPADI	Shoulder Pain and Disability Index
SRM	Standardized response mean
SST	Simple Shoulder Test
ULFI	Upper Limb Functional Index
UW-QoL	University of Washington Quality of Life Questionnaire
UW-QoLv4	University of Washington Quality of Life Questionnaire, version 4
VAS	Visual analog scale
ZSTD	z-standard
α	Cronbach's alpha

Summary of the Chapter

This chapter outlines the framework for the research study through defining the specific research aims and investigational questions to be addressed.

Chapter 2: Review of the Literature

Introduction to the Chapter

Chapter 2 will include an overview of topics related to head and neck cancer (HNC), including a definition, risk factors, incidence and prevalence, and medical management strategies. The chapter will also include a review of the relevant anatomy with emphasis on the Spinal Accessory Nerve (SAN), cervical plexus, and the trapezius muscle, followed by a discussion regarding the implications for HNC-related shoulder dysfunction and rehabilitation. An overview of Classical Test Theory (CTT), item response theory (IRT), and Rasch analysis will then be provided, followed by an extensive review of shoulder-related PROs recommended for quantifying shoulder impairment in the HNC patient population. The chapter will conclude with a summary of what is currently known and unknown about the research topic and the anticipated contributions the study will make to the physical therapy and HNC fields.

An Overview of Head and Neck Cancer

Introduction to the Section

This section will first provide a definition for HNC, and a description of the risk factors, incidence, prevalence, and survival rates of the diagnosis. Common medical management strategies, with a focus on surgical management, will then be discussed.

Definition, Risk Factors, Incidence & Prevalence

HNC is characterized by tumors arising in the upper aerodigestive tract, including the oral cavity, pharynx (nasopharynx, oropharynx, and hypopharynx), larynx, paranasal sinuses and nasal cavity, and the salivary glands. Tumors are most frequently characterized as squamous cell carcinomas (SCCA).[23] Although the presentation and management are often similar to HNC, cancers of the brain, eye, esophagus, thyroid gland, scalp, skin, and muscles and bones of the

head and neck, are not characterized as HNC.[23] Risk factors for HNC predominately include alcohol and tobacco use, and human papillomavirus (HPV) infection related to high-risk sexual behavior.[23] Other risk factors include the use of Paan (betel quid), Maté, poor oral hygiene, occupational exposure, radiation exposure, Epstein-Barr virus infection, and Asian ancestry.[23]

In 2014, HNCs will make up approximately 4% of all cancers diagnosed (42,440 cases) in the United States, while 8,390 people are expected to die from the disease.[24] The current one-year survival rate for HNC is 83%, and the 5-year and 10-year survival rates are 62% and 51%, respectively.[24] Chaturvedi and colleagues suggest that the overall increase in HNC incidence in the United States (56.4% from the 1980s -2000s) is a result of an increasing prevalence of HPV-positive oropharyngeal cancers, and an overall decline in HPV-negative tumors related to alcohol or tobacco use.[25] The increasing prevalence of HPV-positive oropharyngeal cancers is creating a "new" HNC patient that is younger, a nonsmoker, and a nondrinker.[26] Although there is an HPV vaccine, its role for prevention of HNC is yet to be determined.[27] The incidence rate for HNC is twice as high in men than women, and in people over the age of 50.[23,24] People with HPV-positive tumors have an increased survival rate than those with HPV-negative tumors.[25]

Medical Management Strategies

Medical management of HNC varies based on the stage and location of the tumor, and the individual's age and comorbidities; and typically includes a combination of surgery, radiation therapy, and/or systemic therapies.[28,29] Surgical management of HNC may include surgical removal of the primary tumor, elective or therapeutic neck dissection, and reconstructive surgery to remediate structural deficits remaining from tumor removal. Tumors of the head and neck most commonly metastasize to the regional lymph nodes located in the neck. Lymph node status is an indicator of prognosis, and knowledge of this status directs treatment strategies.[30] A

unilateral or bilateral neck dissection procedure is therefore utilized in the presence of known or suspected regional metastasis, and is intended to remove at-risk lymphatic and non-lymphatic structures.[30]

The invasiveness of a neck dissection procedure is dependent on the extent of disease, and is classified based on the levels from which lymph nodes are excised.[30] The American Head and Neck Society and the Committee for Head and Neck Surgery and Oncology of the American Academy of Otolaryngology-Head and Neck Surgery have endorsed a classification system for neck dissection procedures based on the anatomic level and (I-VI) and three sublevels (IA/IB, IIA/IIB, VA/VB) of the neck.[31] A radical neck dissection (RND), first described by Crile in 1906,[32] includes removal of the lymph nodes in levels I to V of the neck, the sternocleidomastoid muscle (SCM), internal jugular vein (IJV), and the SAN. A modified radical, or functional, neck dissection (MRND) removes the lymph nodes in levels I through V, but spares a combination of the SCM, IJV, and SAN. A selective neck dissection (SND) is the least invasive and involves removal of the lymph nodes only in the areas of greatest risk, based on known patterns of metastasis.[30] SNDs are named based on the cervical levels excised.[30,31] An extended neck dissection is more invasive and includes removal of one or more additional lymph node groups or non-lymphatic structures in addition to what is typically removed in a RND.[30] Improved diagnostics and surgical techniques have allowed for the evolution of the RND to MRND and SND, when medically indicated. The surgeons' ability to limit post-operative complications, including shoulder dysfunction resulting from nerve sacrifice or intraoperative damage to the SAN, is one of many positive outcomes of less invasive surgeries.[31,33-41]

An Overview of Relevant Anatomy

Introduction to the Section

This section will provide a review of the relevant anatomy to HNC, including the SAN, the cervical plexus and the trapezius muscle. Anatomical variations, and variations in motor and sensory input for each these structures will also be addressed.

Spinal Accessory Nerve

Cranial nerve XI has two parts, the cranial nerve and the spinal nerve. The cranial nerve, also known as the accessory nerve, originates internally from the nucleus ambiguous and joins with the vagus nerve to innervate the palatal, pharyngeal, and laryngeal muscles. The spinal nerve, or the SAN, originates in the ventral medulla and externally from the upper five cervical segments. The cranial and spinal fibers ascend through the foramen magnum and then exit the skull base through the jugular foramen with the glossopharyngeal nerve and the vagus nerve. The SAN enters the neck in close proximity to the internal carotid artery and IJV, bisecting the anterior triangle of the neck at level II. The nerve passes deep to the SCM where it pierces the muscle, emerging distally to cross the posterior triangle (level V) and enter the anterior border of the trapezius muscle.[30,42-45] Research via intraoperative electromyography (EMG) suggests that the cranial branch of the SAN also descends to the trapezius muscle, providing innervation to the descending (superior) portion of the muscle.[35]

Anatomical variations to the accepted anatomy and anatomical course of the SAN are reported in the literature. These anatomical variations in cervical anatomy from patient to patient and from side to side can result in unintentional intraoperative injury to the SAN.[33,46] For example, a case report published by Bater and colleagues demonstrates a rare anatomical variation in which the SAN divides into two branches proximal to the SCM.[47] In another study,

Lee summarizes variations of the SAN innervation of the SCM, in which the SAN penetrates the SCM in only 54.1% of the cases, sending a branch to innervate the muscle in other cases.[44] Another commonly cited anatomical variation exists in the course the SAN takes in relation to the IJV within the anterior triangle (level II) of the neck. A cadaveric study of 84 necks performed by Saman and colleagues found that the SAN passed anterior to the IJV in 80%, posterior in 19%, and in one case the nerve bifurcated the IJV.[48] Lee and colleagues' findings differed in that they found 39.8% of SANs crossing anterior to the IJV, 57.4% passing posterior, and 2.8% passing through the vessel.[44]

Cervical Plexus

The cervical plexus is comprised of the anterior portions of the first four cervical nerves, which form a series of anastomoses and generate five superficial or cutaneous branches and 10 deep or muscular branches.[49] Although variations in the anatomical organization of the cervical plexus exist, branches of the cervical plexus are thought to innervate the trapezius muscle. According to Pu and colleagues, "C2, C3 and C4 innervation does provide motor input to the trapezius muscle, but the innervation is not consistently present and, when present, does not consistently innervate all three parts of the trapezius muscle."[50(p571)] Lee and colleagues report contributions to the SAN by the cranial nerves at C2 in 53.1% of the cases, C2 and C3 in 38.1% of the cases, and C3 in 8.8% of the cases.[44] In some cases the second and third cervical nerves join with the SAN deep to the SCM prior to entering the trapezius muscle,[42,51] and in other cases, the fibers originating from C3 and C4 cross the posterior triangle independently entering and innervating the middle and lower portions of the muscle.[42,51,52] If the cervical plexus does innervate the trapezius muscle, sacrifice of the cervical root branches of the plexus can result in shoulder pain. Garzaro and colleagues compared individuals with neck dissection with

preservation of the cervical plexus and neck dissection with sacrifice of the cervical plexus and found greater shoulder morbidity in individuals in which the plexus was sacrificed.[53]

Trapezius Muscle

The trapezius is a large muscle that is made up of three parts, the superior/descending portion, the middle/transverse portion, and the inferior/ascending portion. Independently, the superior fibers elevate the scapula, the middle fibers retract the scapula, and the inferior fibers depress the scapula.[54] When working in synergy the three portions stabilize the scapula against the thorax and upwardly rotate the scapula allowing for elevation of the arm, a movement defined as scapulohumeral rhythm.[50,55] The trapezius muscle is most active between 35 and 140° of shoulder abduction, with the greatest force occurring at 90°.[56] At rest, the trapezius muscle provides passive support to the shoulder.[56]

Anatomical variations related to the innervation of the trapezius muscle exist across individuals. It is accepted that the trapezius muscle receives motor innervation from the SAN[50,57] and, in many cases, branches of the cervical plexus.[33,42,51,53] A study of 30 cadaveric necks revealed motor innervation of the trapezius in all specimens by both the SAN and cranial nerves.[51] Some anatomists suggest that the SAN and cervical nerves can also offer sensory (pain and proprioceptive) innervation to the muscle,[50,58-60] although the source of innervation varies based on the study.[50,58] A study performed by Dilber and colleagues, however, claims that while the cervical plexus offers sensory innervation to the neck, it does not offer sensory innervation to the trapezius muscle.[61]

Neurological Implications of the Medical Management for Head and Neck Cancer

Introduction to the Section

This section will discuss the factors that result in peripheral nerve injury to the SAN and cervical plexus ultimately resulting in trapezius muscle atrophy and shoulder dysfunction. Specifically, implications related to surgery, radiation and/or chemotherapy are discussed.

Surgical Implications and Peripheral Nerve Injury

Shoulder pain and weakness secondary to trapezius muscle atrophy may result from intraoperative sacrifice of the SAN in a RND or, in the case of nerve-sparing surgeries in which trauma to the SAN occurs.[1,30,62] EMG studies comparing outcomes for people receiving SND, with little or no trauma to the SAN, shows that those receiving surgical excision of the cervical plexus experienced greater dysfunction of the upper trapezius muscle than those who do not.[53,63] The trapezius innervation from the cranial nerves supports the surgical and clinical experience in which some people maintain shoulder function despite known injury to or sacrifice of the SAN.[34,51,64] Saunders and colleagues therefore recommend that the cervical plexus is also spared, when possible, to minimize shoulder dysfunction.[36]

Intraoperative neurologic injury to the structures innervating the trapezius can occur from traction, compression, skeletonization, thermal injury, and de-vascularization, and are most common at the SAN with dissection or biopsy at levels IIB and V.[30,34,57,63,65,66] The degree of injury and the time to recovery are dependent on the type of nerve injury sustained.[55] There are three classifications of nerve injury - neuropraxia, axonotmesis, and neurotmesis.[67,68] Neuropraxia results from a transient block of the nerve, and will resolve spontaneously within three months. Axonotmesis results from a disruption of the axon and myelin sheath, triggering Wallerian Degeneration and nerve regeneration.[69] Complete denervation, characterized as

neurotmesis, results from complete sacrifice of the nerve, as with a RND. In this case, spontaneous nerve regeneration will not occur.[67,68]

Nerve regeneration occurs at a rate of one inch per month.[67,68] The SAN is 4-5 centimeters in length (cm) when it is lax and 9-10 cm at full length (approximately four inches).[59] In HNC survivors' nerve regeneration and re-innervation of the trapezius muscle can take 6-12 or 15 months or longer to occur.[52,70] Orhan and colleagues studied motor conduction of the SAN in 42 necks using needle EMG studies of the trapezius muscle pre-operatively and post-operatively at three weeks, three months and nine months. The study demonstrates "sufficient re-innervation" at three and nine months when mild partial denervation of the SAN occurred, "moderate re-innervation" for moderate partial denervation, and "mild re-innervation" with severe partial denervation. Full recovery to pre-operative levels did not occur for any subject by nine months, however. Subjects with complete denervation experienced no re-innervation at three or nine months.[71]

Although not related to peripheral nerve injury, surgical reconstruction, such as pectoral cutaneous flap or radial cutaneous flap, may also be a risk factor for increased shoulder dysfunction. Post-operatively patients undergoing surgical reconstruction were found to have a 25° deficit in forward flexion range of motion (ROM) compared to patients not receiving reconstructive surgeries.[72] Similarly, those receiving flap surgery report decreased overall quality of life (QOL) than those not requiring reconstructive surgery.[73] Merve and colleagues did not find a significant difference in shoulder function between patients receiving neck dissection with pectoralis major flap reconstruction and patients receiving neck dissection only 1-1½ years following surgery, however.[74]

Implications Related to Radiotherapy and Chemotherapy

The overall QOL for HNC survivors receiving radiotherapy is worse than those who do not.[73,75] The impact of radiation and chemotherapy on aspects related to QOL, including shoulder pain and function, remains unclear. For example, Short and colleagues found that radiation alone did not cause shoulder pain in 75% of their sample,[76] a finding also supported by Chaplin's research,[77] whereas all patients reporting shoulder pain in Teymorrtash's study had received radiation therapy.[78] When considering shoulder function, in the absence of surgery, radiation and/or chemotherapy do not seem to have a significant effect on shoulder function.[39,65,75,76,79-82] Studies have come to similar conclusions in individuals who also received surgery. For example, Kuntz and Weymuller found no correlation between shoulder dysfunction and the presence or absence of adjuvant radiation therapy at baseline, six and 12 months following surgery.[83] Similar findings are noted for shoulder ROM five years following the conclusion of cancer treatments and radiation.[75] The provision of chemotherapy alone also does not seem to impact patient outcomes with respect to shoulder function.[81]

Other studies suggest increased shoulder-related morbidity in the presence of radiation and/or chemotherapy interventions.[84-86] In 1989, Nowak and colleagues reported radiation therapy combined with neck dissection surgery increased morbidity, including loss of cervical and shoulder ROM.[85] Gane and colleagues report similar findings.[87] In addition, Schuller and colleagues found when radiotherapy and surgery were provided, patients had increased reliance on others for daily activities and decreased participation in social activities.[81] It is possible that the common side effect of radiation fibrosis contributes to these findings.[87] Radiation fibrosis is characterized by progressive tissue fibrosis and sclerosis of all tissues in the radiated field, including skin, muscle, ligament, nerve, and bone. The onset of radiation fibrosis can be

immediate or delayed, occurring greater than three months following the conclusion of therapy. The compressive forces generated by radiation fibrosis, in addition to the ischemia resulting in altered blood supply, can result in neurologic injury leading to neuropathic pain, myelopathies, and plexopathies.[84] Exposure to chemotherapeutic agents enhances the risk of radiation fibrosis.[88]

Head and Neck Cancer and Shoulder Dysfunction

Introduction to the Section

A systematic literature review performed by Goldstein and colleagues provides a comprehensive assessment of shoulder impairments and disability, based on the World Health Organization International Classification of Functioning, Disability and Health (ICF),[89] after neck dissection procedure.[90] This section will define HNC-related shoulder dysfunction, and provide an overview of the risk factors, prevalence, and prognosis for recovery. Factors that contribute to shoulder dysfunction including ROM, strength and pain will be reviewed, followed by a discussion related to the impact of shoulder function on QOL.

Shoulder Dysfunction: Definition, Risk Factors, Prevalence and Prognosis

Shoulder dysfunction, first described by Ewing and Martin in 1952[1] and Nahum and colleagues in 1961,[91] is most commonly caused by sacrifice of or intraoperative trauma to the SAN during a neck dissection procedure.[1,30,62] Individuals with this "shoulder syndrome" typically present with the following postural and functional characteristics: drooping of the affected shoulder, increased prominence of the medial and superior angle of the scapula (scapular winging) both at rest and with movement, shoulder shrug weakness, limitations in active shoulder abduction and flexion with preservation of passive ROM, compensatory

hypertrophy of other muscles with action at the scapula, and pain across the superior border of the trapezius muscle resulting from fatigue of the levator scapula and rhomboids.[1,59,92]

Risk factors for developing shoulder dysfunction following a neck dissection surgery include the location and stage of the tumor, levels dissected, number of nodes removed, and the degree of reconstruction required. The prevalence of shoulder dysfunction is highest in individuals who receive RND because of the intentional sacrifice of the SAN, and ranges from 47-100%.[1,37,40,41,52,72,81,92-96] Shoulder symptoms are less frequently reported in nerve-sparing surgeries. For example, shoulder symptoms are reported in 18-77% of MRND cases,[36,41,50,81,92-94,96] and 15-50% of cases in SND.[50,65,66,78,92-94,96,97] In general, regardless the surgery, dissection occurring at level IIB and/or level V place the SAN at the greatest risk.[73,79,97,98]

Research supports a typical progression of shoulder dysfunction and subsequent recovery based on type of neck dissection received and time from surgery. According to Dijkstra and colleagues, 70% of a sample of patients undergoing RND, MRND, or SND experienced shoulder symptoms prior to discharge from the hospital post-operatively.[72] Subjects receiving a SND typically have returned to baseline shoulder function six months following surgery. Those receiving MRND have minor recovery of function at six months and near full recovery at 12-18 months. Those receiving RND have significant shoulder dysfunction at six months, which does not recover at 12 or 18 months. However, no subject that received MRND or RND regained normal trapezius strength or innervation 12 months following surgery.[39,40] This general progression is supported by multiple studies that specifically assess the impact of HNC treatments on QOL and strength, ROM, and pain at the shoulder joint.

Shoulder dysfunction can persist even five years following the completion of cancer treatments for both SAN-sparing and sacrificing procedures.[75] Although the levator scapula,

rhomboids and serratus anterior can compensate for weakness of the trapezius muscle during arm elevation,[56,99] loss of motor innervation to the trapezius muscle results in resting postural abnormalities associated with downward rotation of the scapula and "sagging" of the shoulder.[56] Long term postural deficits may also result in upper extremity symptoms related to traction of the brachial plexus.[45,60] Other reported outcomes of SAN palsy include subacrominal impingement,[55,60] hypertrophic sternoclavicular joint,[99,100] and secondary adhesive capsulitis.[38,70,99,101] Adhesive capsulitis is a late sign of SAN palsy in that it is often cited as the culprit for persistent joint stiffness and discomfort once the SAN has recovered 12-18 months following surgery.[38]

Shoulder Dysfunction: Clinical Presentation Related to Range of Motion, Strength, Pain, and Quality of Life

Range of Motion

Impaired ROM, specifically shoulder abduction, is a hallmark sign of SAN palsy, although shoulder flexion is also impacted. Impaired shoulder abduction results from the failure of the trapezius muscle to stabilize the scapula sufficiently during glenohumeral joint abduction, resulting in early activation of the deltoid.[93] ROM deficits may be present immediately following surgery and persist for longer than five years.[72,75] For example, post-operative ROM deficits, when compared to the opposite shoulder, have been reported at 47° for abduction and 21° for forward flexion.[72] Individuals receiving nerve-sacrificing surgeries experience greater ROM deficits than those receiving nerve-sparing surgeries.[72,78] Ewing and Martin describe the classic shoulder syndrome in patients undergoing RND as an inability to abduct the arm greater than 90°.[1] Some patients receiving a MRND or RND experience an immediate post-surgical deficit of 55° in abduction compared to those receiving SND.[72] Other authors have reported even greater

restrictions in a similar populations with an average shoulder abduction ROM of 58° (range 30-90°) and average forward flexion ROM to 91° (range 75-120°).[59]

The timeframe for recovery of shoulder ROM outside of the post-operative period varies based upon the reference. For instance, some authors suggest that ROM deficits continue to decline in the post-operative period, and then gradually improve to baseline or equal to controls six to18 months following surgery.[38,39] Another study found that subjects undergoing ipsilateral MRND or bilateral neck dissections were found to have abduction ROM deficits of 50-60° six and 12 months following surgery.[39] There is evidence to suggest that ROM deficits persist years following surgery. Five-year cancer survivors undergoing nerve-sacrificing surgery had flexion and abduction ROM ranging from 100 to 140°, and those receiving nerve-sparing surgeries had flexion and abduction of approximately 140°,[75] however Teymoortash and colleagues report a statistically insignificant difference in shoulder abduction nearly three years following surgery.[78]

In summary, SAN palsy results in limitation of shoulder ROM, most notable in shoulder abduction. ROM deficits are more pronounced in nerve-sacrificing procedures and can be present immediately following surgery. ROM improves with re-innervation of the trapezius muscle and can take 6-18 months to occur, although some studies demonstrate chronic shoulder ROM deficits even five years after surgery.

Strength

Muscle strength is a contributing factor to shoulder function, and can be a result of SAN-palsy related denervation of the trapezius, and overall loss of lean muscle mass occurring during cancer treatments. This section will highlight published research related to shoulder strength following various neck dissection procedures, in addition to the available literature related to overall deconditioning and muscle atrophy as a mechanism for loss of strength.

Isokinetic shoulder muscle strength testing performed by Cheng and colleagues[93] one and six months post-operatively highlights the variations in shoulder strength following RND, MRND, and SND. Patients receiving a RND demonstrate statistically significant and lasting strength deficits compared to pre-operative and contralateral measurements for shoulder flexion-extension, abduction-adduction, and internal-external rotation at one and six months following surgery. Patients receiving MRND also experienced statistically significant strength deficits with the same movement patterns at one month however had some improvements at six months. In patients receiving SND flexion-extension and abduction-adduction movement patterns were significantly weaker at one month but had returned to pre-operative levels by six months. These findings were supported by EMG studies of the trapezius muscle at five weeks post-surgery, which demonstrated abnormal EMG findings and diffuse denervation on the operated side in RND. The EMG findings in the MRND group were significantly lower than the contralateral shoulder, and there were no abnormal findings in the MRND and SND groups.[93] Remmler and colleagues report similar findings with manual muscle testing of the upper and middle trapezius. Upper and middle trapezius strength declined in the post-operative period for both nerve-sparing and nerve-sacrificing procedures. While strength returned at six and 12 months to near normal preoperative levels in the nerve-sparing group, strength remained significantly lower than normal in the nerve-sacrificing group.[40] Saunders and colleagues report 67% of patients receiving RND and 47% of patients undergoing MRND continued to demonstrate signs of muscle atrophy of the trapezius at a mean follow-up time of six years post-surgery.[36] Another study of 92 neck dissections (MRND/SND and RND) found greater ROM and strength (elevation, abduction) deficits at the glenohumeral joint in those receiving RND however both groups demonstrated

greater deficits than the control group six months after surgery suggesting some degree of SAN impairment in all cases.[80]

Individuals undergoing any combination of treatment for HNC experience significant weight loss during treatments, 72% of which is attributed to loss of lean muscle.[102,103] In individuals undergoing combined chemo radiotherapy, weight loss can begin as early as the first week of treatment.[103] This significant loss of weight is a result of decreased caloric intake,[102] and a decline in physical performance and increased functional dependence on caregivers.[103] In general, patients with HNC fail to meet recommended daily exercise levels with the majority being completely sedentary.[104] Commonly cited barriers for physical activity include dry mouth or throat, fatigue, drainage in the mouth or throat, difficulty eating, shortness of breath, and muscle weakness.[105] Decline in lean muscle mass and function may contribute to the persistent strength deficits and disability of the upper quarter. For instance, a study performed by McGarvey and colleagues found the expected strength deficits of the upper and middle trapezius at the ipsilateral shoulder following neck dissection surgery, however also found that the contralateral shoulder has significant strength deficits compared to healthy controls.[106]

Pain

There are inconsistencies in the prevalence of shoulder pain in the setting of SAN palsy. Cheng and colleagues report 57% of patients undergoing any neck dissection report shoulder pain,[93] whereas reports by Van Wilgen and Dijkstra cite 69-70% experience shoulder pain after neck dissection.[62,72] Reported incidence rates are lower in a study performed by Chaplin and colleagues, in which 14% of patients reported shoulder pain at diagnosis, 37% at 12 months, and 26% at 24 months. Of these patients, only 1-5% characterized their pain as severe during the two-year timeframe.[77] The incidence of pain following RND is higher than with less invasive

neck dissection surgeries, with 42-100% of patients reporting pain.[1,72,93,107] There is a significantly decreased incidence of reported shoulder pain in nerve-sparing procedures, with 39% of those receiving MRND[76] and 5.8% of those receiving SND reporting shoulder pain.[78] Chaplin, however, did not find a significant difference in prevalence or severity of shoulder pain based on neck dissection procedure.[77]

Quality of Life

Shoulder dysfunction is a well-documented side effect of neck dissection surgery and a significant prognostic indicator of QOL.[75,77,98,108-110] Shoulder dysfunction in the setting of SAN palsy results in decreased ability to perform basic activities of daily living (ADLs), recreation and employment activities, all of which contribute to QOL. Specifically, individuals report difficulty with overhead activities, including putting on a shirt, turning on a light switch, or combing hair, heavy lifting, and cutting food, prolonged writing and driving a car.[56,60,109] As such, QOL and function seem to be less when the neck dissection side matches hand dominance.[111]

Descriptive studies assessing differences in QOL in patients receiving SND, MRND, and RND mirror objective findings of ROM and strength deficits. Subjects receiving SND report near normal function on the University of Washington Quality of Life Questionnaire (UW-QoL) at six months, while those receiving MRND and RND report persistent shoulder dysfunction and decreased QOL. By 12-18 months, those receiving MRND have recovered function similar to the SND however QOL scores remain lower than pretreatment levels. Those receiving RND continue to report increased shoulder dysfunction and decreased QOL at 12 months.[73,83] Reported QOL in HNC survivors five years following both nerve-sacrificing and sparing surgeries remains impacted by the presence of shoulder dysfunction.[75] Interestingly, one study

found that people with HNC tend to demonstrate improvements in subjective reports of shoulder function and/or QOL over time despite persistent objective findings of persistent shoulder dysfunction.[98]

Schuller and colleagues studied employment-related disability in a population of patients receiving MRND and RND procedures and found that only 51% of patients returned to their prior employment following cancer treatments for HNC. There was no significant difference in return to work between surgeries. However, there is a relationship between the degree of physical activity that is required in the job and the rate of return, with 38% returning to work at a job classified as "very strenuous" and 54% returning to a job deemed "not strenuous."[81]

In summary, shoulder dysfunction is a common side effect of neck dissection and SAN palsy and is most prevalent in neck dissections that sacrifice the SAN. Shoulder dysfunction is characterized in the literature as shoulder pain, and limitation of ROM and strength. These impairments result in decreased ability to perform daily functional, recreational and employment activities – all components of the construct of QOL. As such, patient-reported QOL is also lower in HNC survivors. Shoulder dysfunction and QOL tend to improve 6-18 months following surgery; however, can persist for five years or longer.

Head and Neck Cancer & Shoulder Dysfunction: The Role of Physical Therapy

Introduction to the Section

Physical therapy can play a valuable role in the recovery of a HNC survivor during and following cancer treatments. This section will provide an overview of physical utilization, available research related to physical therapy interventions and efficacy, physical activity trends in the HNC population, and a brief introduction to importance of patient-reported outcome measures (PRO) to quantify impairment and show value of physical therapy interventions

provided. A more thorough description of available shoulder-related PRO will be discussed in later sections.

Trends in Physical Therapy Utilization

It is postulated that exercise interventions can improve reported QOL of HNC survivors by improving both cervical and shoulder function,[110] nonetheless, there is evidence to suggest that referral to physical therapy is underutilized in this population.[75,112,113] Often, physical therapy referral does not occur until adhesive capsulitis and subsequent disability have occurred.[113] In a sample of 105 HNC survivors, only 15% report receiving exercises for shoulder dysfunction from a physical or occupational therapist. Six percent received exercises from a physician or a nurse, and 6% received exercises from both a medical and therapy provider.[75] Despite low referral rates, authors frequently recommend referral to physical therapy for shoulder dysfunction following neck dissection surgery for restoration and/or maintenance of ROM and strength of the neck and shoulder, strengthening of the scapular stabilizers, teaching joint protection and compensatory strategies, and decreasing pain.[99,109,113-116]

Shoulder Rehabilitation: A Review of the Literature

Despite the frequent support of physical therapy in the literature, the effectiveness of physical therapy in the HNC patient population has yet to be definitively proven. This section will provide a detailed summary of the literature available related to physical therapy management of shoulder dysfunction in the setting of HNC. Finally, the physical therapy management of individuals who have received muscle or nerve grafting surgical procedures to optimize shoulder function is mentioned.

According to Kuntz and Weymuller, outcomes with respect to shoulder function are dependent on the "degree of injury to the SAN intraoperatively, variations in innervation of the

trapezius muscle, and the amount of physical therapy received postoperatively."[83(p 1337)] Interestingly, subjects enrolled in the study performed by Kuntz and Weymuller were not enrolled in physical therapy, and were instead instructed to pursue non-specific "strengthening and ROM exercises at home" with no formalized instruction provided.[83] Watkins and colleagues suggest that instruction in a home exercise program that includes stretching exercises before discharge from the hospital is just as effective as physical therapy, however he cites limitations in his study design and suggests that referral to physical therapy postoperatively "will save both the patient and the doctor possible morbidity."[65(p 618)]

In 2010, McGarvey and colleagues published a literature review, which systematically reviewed and critically appraised research related to HNC-related shoulder dysfunction and physical therapy through 2009.[117] The authors retrieved 20 articles from the literature, and excluded 11 based on low levels of evidence (Sackett's level of evidence four and five).[118] Nine articles remained for review, five of which were pre- and post- study designs (level of evidence: three).[40,119-122] From their review, McGarvey and colleagues concluded that although there is some evidence to support an indirect effect of physical therapy in patients with a potential for nerve recovery, there is a lack of evidence to support the effectiveness of, and the timing and type of physical therapy needed to address shoulder dysfunction in the HNC patient population.[117] The authors suggest several reasons for the lack of robust clinical trials assessing the effectiveness of physical therapy in the HNC patient population including the ethical issues associated with withholding physical therapy from patients and the tendency for patients to be receiving adjuvant cancer-related treatments which would compound study results and decrease enrollment rates.[117]

In 1987, Herring and colleagues published a description of a rehabilitation program for patients following a RND procedure. The exercise program progressed the patient through a series of passive, active assisted and active ROM exercises in gravity eliminated progressing to anti-gravity positions. The authors also included isokinetic strength training using a Cybex machine with progression to a home-based exercise program including ROM and isotonic shoulder strengthening for varying planes of shoulder elevation to 90°.[56] Although this paper was not experimental in nature, it is valuable in that it provides a thorough description of exercise progression for this patient population. The five pre- and post-study design studies included in the McGarvey review highlight the variation and progression of physical therapy intervention over the past four decades.[40,119-122] An early descriptive study published in 1978 detailing physical therapy-supervised home exercise program for shoulder dysfunction in the setting of SAN sacrifice utilized infrared luminous lamps, strengthening of the scapular stabilizers and elevators, ROM exercises, and stretching to the serratus anterior muscle. The authors claimed that the physical therapy treatment provided significant or total pain relief in all 16 patients and postural improvements in 15 patients.[120] A similar physical therapy intervention was also utilized in a manuscript published by the same authors in 1975.[99] In another study published in 1986, all participants received physical therapy aimed at maintaining ROM of the shoulder joint. The results of the study were not focused on the outcomes associated with physical therapy intervention, rather the variations in shoulder function based on surgery type.[40] A 1988 study utilized physical therapy intervention which included "constant (electrical) current, exponential and progressive current for the trapezius, massage for the neck and shoulders, and therapeutic exercises to the affected shoulder."[121(p 144)] While the type of therapeutic exercise was not described, the authors concluded that physical therapy was beneficial in improving the

stabilization of the scapula through strengthening of the serratus anterior, rhomboids, and levator scapula, decreasing pain, atrophy, stiffness and pathological changes at the shoulder.[121]

Studies have also addressed the role of shoulder orthosis in management of shoulder dysfunction.[119,122] Kizilay and colleagues performed a study in which 34 subjects wore the orthosis with effects tested at baseline, and one, three, six and 12 months postoperatively. The authors report positive outcomes with use of the orthosis,[122] however the lack of a control group limits the ability to claim that the orthosis caused the outcome.

The remaining four studies included in McGarvey's literature review[117] include a prospective study (evidence level three),[123] a retrospective cohort study (level three),[124] a pilot for a randomized controlled trial (RCT) (level two),[114] and a RCT (evidence level one).[115] The prospective study published by Salerno and colleagues in 2002 assessed the efficacy of rehabilitation on pain and dysfunction in a sample of 60 patients following a MRND and total laryngectomy. The study utilized a two-group prospective design. Subjects were selected to receive physical therapy if they lived in close proximity to the hospital, and were otherwise assigned to the control group. The authors utilized interventions aimed entirely at restoring passive ROM of the shoulder in the post-operative stage (range 30 – 180 days) under the premise that the achievement of full passive shoulder ROM will result in the spontaneous recovery of active mobility. The authors found that those who received physical therapy experienced decreased shoulder dysfunction and improved QOL compared to those who did not six months following surgery.[123]

The retrospective study is a Japanese study published in 2007, which studied the role of occupational therapy on shoulder pain, ROM and QOL following RND. The authors compared an occupational therapy group of 35 shoulders to a no-therapy group of 10 shoulders. The

occupational therapy group initiated therapy on average 55 days (range, 11-263 days) from RND and received services five days a week throughout their hospital stay (average 74 days). The no therapy group was on average 131 days from RND (range, 70-2466 days). Occupational therapy included muscle relaxation techniques, massage, ROM (passive, active-assisted) exercises, and ADL training. The results of the study showed that occupational therapy did not have an impact on shoulder pain in the setting of SAN palsy, but did improve shoulder ROM and return to independence in ADLs and housekeeping activities.[124] Similar findings were reported in a 10-patient case series by the same group of authors in 2002.[125]

In 2004, McNeely and colleagues published a pilot study for a RCT that compared standard care to a 12-week progressive resistance exercise training (PRET) program. The pilot study considered subject willingness to participate in a 12-week PRET program for shoulder dysfunction in the setting of SAN neuropraxia or neurectomy following RND, MRND or SND surgery, and the effects of physical therapy intervention on shoulder function, pain, disability and QOL. Subjects less than eight weeks from surgery, in some cases currently undergoing radiation therapy, and subjects greater than eight weeks from surgery were compared for study and exercise adherence. The PRET group performed a series of ROM and stretching exercises and progressive strengthening exercises for the rhomboids, levator scapula, biceps, triceps, rotator cuff, and deltoid; and the standard of care group performed a home based program of ROM, stretching, and strengthening exercises of the rhomboids and levator scapula. The PRET group demonstrated statistically significant improvement in shoulder external rotation active ROM and Shoulder Pain and Disability Index (SPADI) scores. No significant difference was found in flexion or abduction ROM or in reported QOL. The authors also demonstrated high study and exercise adherence rates.[114]

In a follow-up RCT published in 2008, McNeely and colleagues compared a PRET program to a standardized therapeutic exercise protocol in patients who had undergone a RND, MRND or SND procedure. Eligibility required that subjects had completed adjuvant treatments including radiotherapy at the time of enrollment. Both groups underwent 2-3 supervised exercise sessions per week for 12 weeks. The standardized exercise protocol included ROM and stretching exercises, postural exercise, and basic strengthening exercises with light weights and elastic bands for various muscle groups. The PRET group received the same ROM and stretching exercises, however the strengthening exercises were replaced with a standardized PRET program. Similar to the pilot study, the RCT showed superior outcomes in the PRET group for the SPADI, shoulder strength and muscular endurance, and active external rotation ROM. Passive abduction ROM was also improved. QOL changes showed a trend toward improvement, but were not statistically significant.[115] In 2015, the McNeely group published a follow up study, which reported maintained patient-reported shoulder function at 12 months following surgery. In addition those individuals who continued resistance exercise training reported better neck dissection-related function on the NDII and shoulder-related QOL on the SPADI than those who did not.[126]

A Cochrane review published in 2012 by Carvalho, Vital and Soares[127] included only three studies that met the rigorous criteria for conclusion.[114-116] The McNeely pilot[114] and RCT[115] included in the McGarvey[117] literature review were included, as was a newer paper published by Lauchlan[116] in 2011. The Lauchlan study utilized a RCT design to compare routine postoperative physical therapy while the patient was hospitalized, including respiratory care and advice for active movement of the neck and shoulder, to an intervention group which received the same routine postoperative physical therapy in addition to outpatient physical therapy for the first three

months following surgery. Outpatient physical therapy intervention included individualized exercise prescription based on patient presentation, and included active ROM, stretching and strengthening of the shoulder, postural care, and neuromuscular re-education of the scapulothoracic postural muscles. The results did not show a significant difference between the intervention and control group with respect to shoulder function or QOL one year following surgery.[116] This study has several design flaws and was classified by the Cochrane review as has having a high risk for bias. The McNeely pilot study was also classified as having a high risk for bias, while the RCT was classified as having a low risk for bias. Pooled data from the three studies (n = 104) suggests that PRET is more effective than standard physical therapy treatment for the management of shoulder dysfunction in the setting of HNC. Neither demonstrate a significant improvement in reported QOL, however.[127] A limitation of the Cochrane review is the heterogeneity of samples included. The Lauchlan study enrolled subjects immediately following surgery, whereas the McNeely studies included subjects who were 2-180 months from surgery. This variation leads to bias in interpretation because of the change in shoulder function that occurs as trapezius muscle re-innervation or compensatory strengthening occurs.

In 2015, McGarvey and colleagues published the results of a single-blind prospective RCT which compared a supervised physical therapy program to usual care at three medical facilities in Australia. Patients undergoing SND or MRND with SAN-related shoulder impairment were randomized to an intervention group, which received progressive scapular stabilization exercises among other interventions, or a control group which received general advice and a brochure of general shoulder and neck exercises. Outcomes, including shoulder flexion and abduction ROM, the Neck Disability Impairment Index (NDII) and the SPADI, were measured at baseline, three months, six months and 12 months. The authors found a statistically

significant difference in shoulder abduction between groups at three months, but no significant difference for shoulder pain, function, or QOL on the NDII or SPADI at any time point. The authors do cite several potential limitations in their study which could have impacted their outcomes, including a high attrition rate decreasing statistical power at the six and 12- month time points and the potential contamination bias of the control group through allowance of individuals (26% of sample) to obtain physical therapy if they desired.[128]

Gallagher and colleagues also published a study in 2015, which showed positive effects of physical therapy as measured by the Constant's Shoulder Score (CS) and NDII. Physical therapy intervention was not the primary outcome of this study, therefore authors did not collect data related to the physical therapy received, duration, frequency or intensity.[129]

Recent studies have considered the effects of alternative therapies including acupuncture[130] and Tai Chi Qigong.[131] Six months of Qigong helped to maintain shoulder function but did not improve it.[131] On the other hand, weekly acupuncture treatments were found to improve shoulder function, as measured by the CS, and shoulder-related QOL, as measured by the NDII, more than five weeks of standard physical therapy. The authors of this study recommend consideration of a combination of PRET-based physical therapy and acupuncture, rather than standard physical therapy alone.[130]

Literature on surgical repair techniques for trapezius paralysis, such as the Eden-Lange dynamic muscle transfer technique, tends to report that conservative rehabilitation is ineffective.[55,60] The Eden-Lange procedure involves the transfer of the levator scapula, rhomboid major and rhomboid minor muscles in a way that their new muscular insertions recreate the force vectors previously offered by the three parts of the trapezius muscle.[55] Bigliani describes conservative management as "physical therapy, transcutaneous nerve stimulation, external

support, chiropractic consultation, management in pain clinic, and use of non-steroidal anti-inflammatory medications as well as narcotic analgesics."[60(p 1535)] It should be noted that the details of the physical therapy provided are not outlined. In addition, the publications used to substantiate the claim that conservative management is ineffective range in publication date from the 1950's to the 1980's, and appear to be based on clinician expertise rather than research.[55,132,133]

Various surgical techniques aimed at restoring shoulder function through nerve grafting are also described in the literature. A study published by Clinton and colleagues describes a single patient case study detailing physical therapy intervention directed to a manual laborer following a MRND with sacrifice of the SAN and subsequent reconstruction using the greater auricular nerve. Physical therapy intervention began six months following surgery and included patient education, scapular taping, soft tissue mobilization, and PRET. Strengthening emphasized functional strength allowing the patient to return to work successfully, however trapezius muscle strength had not recovered 12 months following surgery.[134]

Physical Activity & the Head and Neck Cancer Survivor

Physical activity patterns in the HNC population are problematic. According to Rogers and colleagues, only 30.5% of individuals with HNC are meeting recommended exercise guidelines (150 minutes of physical activity per week and strength training twice a week) at the time of diagnosis. Physical activity declines further after diagnosis with only 8.5% of individuals meeting recommended exercise guidelines. Rogers and colleagues demonstrate a significant association between weekly physical activity, functional well-being, overall QOL and fatigue[104]; and Fong and colleagues report improved cervical and temporomandibular joint mobility, and

sleep quality in Cantonese patients with nasopharyngeal cancers who participated in a 6-month Tai Chi Qigong training program.[131]

The already low physical activity levels in this population make adherence to exercise prescription following diagnosis difficult. Factors that may contribute to decreased engagement in physical activity include xerostomia, fatigue, swallowing difficulties, shortness of breath, pain, communication difficulties, and dissatisfaction with QOL.[105] McNeely and colleagues suggest that more extensive, nerve-sacrificing surgeries and alcohol consumption are predictors for low exercise adherence rates.[135] Rogers and colleagues offer additional factors to non-adherence including increased levels of fatigue during the sixth week of radiation therapy and lasting for several weeks following the conclusion of treatment, the patient's desire to put the cancer diagnosis and treatment behind them once the cancer treatments have ended (including exercise prescribed during cancer treatments), depression and anxiety.[136]

Quantifying Impairment and Demonstrating Value

Physical therapy referral at diagnosis may be of benefit to the patient. Pre-operative assessment can uncover preexisting postural, ROM, and/or strength deficits, provide appropriate exercises to address these deficits in order to minimize post-surgical changes, and provide education on posture, joint protection and positioning. Establishment of baseline shoulder function and/or QOL through validated questionnaires would also be of benefit.[137]

The Centers for Medicare and Medicaid Services (CMS) have mandated the use of functional outcome reporting in the assessment and management of patients by physical, occupational, and speech therapists.[22] A future shift to pay for performance reimbursement models will also require use of objective measures to quantify change related to interventions provided. PROs defined as "instruments that elicit the individual's observation of his or her

experience,"[138(p 193)] are commonly used to satisfy CMS mandates, to demonstrate treatment effect in the clinical setting, and as a measured variable in research. PROs can be general assessments of health-related QOL (HRQOL), or specific to a disease or diagnosis.[138] Traditionally, there are numerous considerations when choosing an outcome measure, including: (1) the diagnosis, age and level of disability of the patient or population of interest; (2) whether short or long-term conditions are of concern; (3) the level of detail required by the assessment; (4) that the test measures what it is intended to measure; (5) the scoring is clear; (5) the measure is responsive to change; and (6) evidence exists for the reliability and validity of the measure.[139] Other considerations include the mode of administration, response option format, recall period, responder burden, translation or cultural adaptation availability, accessibility to the test, scoring and interpretation information, and the presence of floor and/or ceiling effects.[138,140]

In summary, while physical therapy referral is often recommended in the literature, HNC survivors are not receiving physical therapy services at an optimal rate or time within their recovery. There are very few high-quality physical therapy-related research studies available for this patient population, leaving clinicians with very little evidence to base clinical decision-making on. Physical therapy research for shoulder dysfunction in the setting of trapezius muscle weakness has demonstrated overall improvements in shoulder function and QOL when compared to those who do not receive physical therapy. In the clinic, physical therapists should use validated PROs to quantify impairment and demonstrate value.

Psychometric Theory Models: Classical Test Theory, Item Response Theory, Rasch Analysis

Introduction to the Section

Psychometric theory, or test theory, is a "theoretically oriented field of study, which is of general relevance regardless of the particular test, scale or measuring instrument used in a given

situation."[141(p 9)] It refers to the statistical and mathematical methods underlying the construction, development, and revision of measurement instruments and their application. Originating in the behavioral, social, and educational fields, the purpose of test development and administration is to quantify various constructs of interest.[141] A construct is an indirectly observable or measurable trait or behavior that exists on a continuum within individuals. Constructs, often referred to as latent variables, are measured by a sampling of content indicators or proxies (test items) that represent various aspects of the construct. While impossible to include all characteristics that make up a construct, a test should provide a good representation of the latent variable and should demonstrate an accurate relationship with other similar, previously established constructs, a concept known as construct validity.[141-143] Because the construct cannot be directly measured, it must be assumed that error exists within the test score.[143]

Two theoretical approaches for creating and evaluating tests are CTT and IRT.[142] The five scales addressed in this research study were developed and have been studied using CTT, which will be described here. IRT offers a contemporary method of psychometric testing which provides a more in-depth analysis of the construct validity and reliability of measures. This section will also provide a description of IRT, in addition to a detailed review of Rasch analysis – a theoretical framework derived from IRT.

Classical Test Theory

CTT is a theoretical framework that analyzes how successful proxy indicators are at quantifying latent variables of interest.[143] CTT is based on the premise that an observed test score (X) is comprised of an individual's true score (T) and measurement error (E), $X = T + E$.[142-145] CTT also assumes that the random error around the true score is normally distributed and, over an infinite number of trials, would equal zero. It is also assumed that random error is

uncorrelated with other random error and with the true score.[146] There are two sources of measurement error, systematic and random. Systematic error occurs in a consistent and repeatable manner and is unrelated to the construct being measured.[147] Random error "represents the combined effect of particular, transient, unrepeatable, and nonsystematic factors…that are unrelated to the attribute of concern."[147(p 116)] Random error may be a result of external and internal factors such as mood or current physiologic state or a factor of uninformed guessing. Both random and systematic errors are impossible to separate from the true score because the true score is a latent construct that cannot be directly measured. As mentioned previously, CTT assumes that random error across multiple trials is zero. The same assumption does not exist for systematic error because it is consistent and repeatable across trials. Systematic error is absorbed into the true score and therefore poses a significant threat to validity of the test.[147] If the test is perfectly measuring the construct of interest, there would be no error and the observed test score and the true score would be equal. Reliability coefficients, expressed as a ratio of true test score variance to observed test score variance, provide an estimate of the accuracy of measurement and the degree of consistency between the observed test score and true score.[142,148] Error in a sample decreases the reliability, or precision, of a measurement. Therefore, the greater the error variance the more unlikely the test will be able to capture change in ability level.[149]

Reliability

Commonly utilized scale characteristics under the CTT framework for reliability are based upon several assumptions for parallel forms described by DeVellis[143] and include temporal stability (test-retest reliability), interrater reliability, internal consistency reliability (Cronbach's alpha), and standard error of the measure (SEM).[138,142,143,150] These characteristics are established for sample-based data, and are therefore influenced by group characteristics. Clinicians and

researchers should therefore choose measures that have been tested in populations similar to the population of interest before generalizing findings.[150,151] Increasing the number of test items is a way to improve reliability; in fact Cronbach's alpha is based upon the correlation between the test items and the number of items within the measure.[143] SEM, an index of the quality of a measurement, is defined as the "average, across persons, standard deviations of the individual propensity distributions on a measure under consideration."[148(p 144)] The SEM is commonly used to construct confidence intervals (CI) around scores.[151] Approximately two-thirds of subjects' observed scores will fall within ±1 SEM of their true score and approximately 95% will fall within ±2 SEM. This is based on the assumption of normal distribution and identical standard deviations (SD) of the individual propensity distributions (true test scores).[148] A problem with SEM is that it incorrectly assumes that test item precision is equal for mid-range and extreme scores.[151]

Validity

Measurement validity, the idea that the test is measuring what it is intended to measure, is also included within the CTT framework. The question whether a measure is valid can never be affirmatively answered, because there are varying degrees of validity. The goal is to establish a strong case for validity through accumulating evidence to more than one of the following primary types of validity: criterion-related, content and construct validity. Criterion-related validity refers to the ability to predict an individual's test score on one measure based on his or her performance on another measure. There are two kinds of criterion validity: predictive and concurrent validity. Content validity refers to the degree that the test items within a measure represent the domain assessed and is therefore also dependent on the operational definition provided for the construct. Content validation often occurs through expert review of items from

content experts not involved in the development of the measure. Construct validity is an all-encompassing form of validity, which also includes criterion-related and content validity, that supports that a test adequately measures the construct of interest.[152] Baghaei defines construct validity as the "trustworthiness of score meaning and its interpretation."[153] Under CTT, construct validity may be established through correlational analysis, differentiation between groups, factor analysis and latent variable modeling, and multi-trait, multi-method procedures supporting convergent and discriminant validity of the measure.[152]

Strengths and Limitations of CTT

CTT is based upon assumptions that are easily met by most data sets.[145,154] It has therefore been widely utilized across various fields of study, including physical therapy, for test development and score analysis.[154] CTT has remained popular over the years for the following reasons: (1) widespread familiarity with CTT terminology and methodology across fields of study; (2) most currently available measurement tools are based on CTT; and (3) easy accessibility to statistical packages that run CTT-based analyses. In addition, CTT parameters often strongly correlate with other modern measurement theory parameters, suggesting that the models are comparable.[143]

Despite its frequent use, CTT has several reported weaknesses related to sample dependency, assumption of test item equivalence, and assumption that SEM is equal across difficulty levels despite ability level. CTT assumes item equivalence, or that each test item contributes equally to the test score, irrespective of the item difficulty level or person ability level. This assumption leads to incorrect test score interpretation when one assumes that a raw score of 50 indicates twice as much disability as a raw score of 25 for example. Similarly, one cannot assume that the difference between a response option of 'agree' and 'strongly agree' is

equal to the difference between 'disagree' and 'strongly disagree.' The SEM is also assumed to be equal at each point along the scale or for every individual. This assumption is incorrect because, by definition, in a normally distributed sample the SEM increases further away from the mean, and is not equal at each point along the scale.[141,142,144(p 299-301)] Other reported weaknesses include: (1) failure to capture individuals with low and high ability levels; (2) inability to equate ability levels across two or more tests; (3) reliance on the idea of parallel forms (respondents remain the same over test trials) for reliability testing; (4) interpretability is based on overall test score and does not take into account performance on individual test items; (5) failure to provide solutions to common testing problems related to designing tests and identifying biased test items; (6) the use of nominal or ordinal level data in mathematical functions and comparisons; and (7) inability to estimate item difficulty and person ability separately.[154-157]

Item Response Theory

IRT, also described as "item response theory," "item characteristic curve theory," or "modern test theory," dates back to 1916 when Binet and Simon plotted performance levels against another variable for test development. However, the "birth" of IRT is attributed to the work of Frederic Lord in the 1960s. The development of IRT is described in detail by Hambleton and Swaminathan.[154] IRT assumes that an individual's performance on a test can be predicted based on the identification of latent traits or abilities, estimating ability scores, and subsequently using those scores to predict performance.[154,158] IRT models assume that the more able a person, the more likely he or she will be to answer higher difficulty items correctly.[142,156] Ability level does not change based upon the test taken, however a person's test score may change based on a test that contains easy or more difficult items.[145]

Strengths and Limitations of IRT

IRT provides a solution for many of the weaknesses reported in CTT. Unlike CTT, IRT allows for score estimation independent of the sample from which the score was obtained, and the ability to equate one measure to another based on the completion of a set of common items or test items. In addition, the scores are transformed from nominal or ordinal level data to interval level data (expressed on a linear scale as a log odds unit, or logit) allowing for the performance of mathematical functions and comparisons (conjoint additivity).[151,159] Data transformation to interval level data solves a problem inherent in CTT related to processes for handling missing data. For example, one method for handling missing items in CTT is to calculate the mean of the sum of answered items, inputting the value in the data set for the missing test item. In doing so, the researcher makes a critical error in using numerical labels from an ordinal scale to perform mathematical functions that require continuous level data. This value is then utilized for data analysis and increases the risk of making a Type I error. This method violates the principle of conjoint additivity. Of note, IRT does not require that each item is answered for analysis to occur.[159]

IRT provides information related to how single test items and persons function within the entire measure or population (item or person fit), dimensionality, floor and ceiling effects, and gaps and redundancies in test item content.[144(p 324)] Information related to gaps and redundancies in a measure allows for modification or development of measures that are shorter in length and more precise. Because each person and item are estimated (referred to as "measures"), IRT models provide an ability to assess the measurement precision for items along the latent construct. Each person and item measure is accompanied with its associated error, called "standard error" (SE). SE is analogous to the SEM in CTT however more accurate than the SEM

because it is based on each person or item measure rather than the entire test score.[142,151] The SE for each individual item can be summed to produce an average error for the scale. This method is more accurate than reporting an error variance for the average person sampled.[151] Another benefit of IRT models is that specific item contribution to the precision of the overall test can be determined. In test development, this information can be utilized to add or remove test items that are found to be imprecise across ability levels.[160]

Assumptions of IRT

IRT is a framework of numerous mathematical models, which are each based on specific assumptions about the data set including unidimensionality or multidimensionality, linear or non-linear models, and dichotomous or multichotomous response options.[154] Multidimensional models however are not fully developed and therefore are infrequently utilized.[146] IRT replaces the "soft assumptions" of CTT with two "hard assumptions" - local independence and unidimensionality.[142,144(p 301-2)] Local independence refers to the idea that individual test items are independent of each other and unrelated to any other item on the test.[142,144(p 301-2)] The lack of local independence is considered statistical dependence.[161] Unidimensionality refers to the idea that "one trait can be used to explain the lack of statistical dependence among" test items.[155(p 274)] The presence of unidimensionality under IRT allows for test results to be equated to another due to the assumption that the test is measuring only one trait, not several.[154] As such, one source suggested a third assumption of IRT - that tests are not administered under timed conditions. A timed test would be a violation of unidimensionality because one would be unable to determine whether a test item was answered incorrectly due to ability level, or because they failed to reach the test item or rushed to answer it.[146]

Hambelton and Jones suggest that IRT models are robust to deviations from these assumptions.[145] Wright and Linacre suggest that it is impossible to remove all latent variables from a construct, leaving an entirely unidimensional scale; however "the pursuit of approximate unidimensionality" is an important endeavor to pursue in order to have accurate interpretation of a total score.[156] If these assumptions are met, one can begin to predict the probability of a correct response on a test item based on the relationship between a test item's difficulty and the person's ability level, which can be graphically depicted in an item characteristics curve.[144(p 301-2)] Ability level across the latent construct is displayed on the horizontal axis, expressed in standard deviations from the a mean ability level of zero, and the probability of answering the item correctly (0-1.0) is on the vertical axis.[162,163]

Models of IRT

The three commonly used models of IRT, the one-, two-, and three-parameter models, vary based on the parameters considered in the model.[164] The three parameters considered in IRT include item difficulty, item discrimination, and guessing. The item difficulty parameter assumes that when all else is equal, a more difficult item will have a greater probability of being answered incorrectly. The item discrimination parameter represents the ability of a test item to discriminate between ability levels. More difficult items are better at discriminating between ability levels than easy items, and are therefore more desirable. The guessing parameter takes the likelihood that test-takers will guess on test items into account, and indicates the probability that an individual of low ability will answer an item correctly based on their ability to guess correctly.[162] This section will provide an overview of each of these models.

One-parameter Model

The one-parameter model, sometimes referred to as the Rasch model, is the simplest of the three models. It assumes that all test items discriminate person ability equally and that no guessing occurs. The only factor considered is item difficulty. Therefore, the item characteristic curves each have the same slope, but are placed at different levels of difficulty along the construct continuum. The one-parameter model also assumes that at the lowest ability level of a trait, the responder has a zero percent likelihood of answering the easiest test item correctly. Therefore, the construct continuum begins at zero (the x-axis on the item characteristic curve.[144]

Two-parameter Model

The two-parameter model allows the item characteristic curves for each test item to vary on two parameters, item difficulty and item discrimination; therefore the item characteristic curves vary both on slope and spacing along the construct continuum.[144] The addition of the item discrimination parameter provides a better understanding of how two test items separate individuals based upon their ability levels.[162] The varying slopes of the item characteristic curves offer a possibility that at some point the curves may cross, suggesting the chance that a low difficulty test item may be answered incorrectly, while a more difficult test item is answered correctly. Like the one-parameter model, the two-parameter model also assumes that an individual with the lowest ability level of a trait will have a zero percent likelihood of answering an item correctly, and the curve therefore starts at zero (x-axis).[144]

Three-parameter Model

The three-parameter model varies from on the one- and two-parameter models because, while it also considers item difficulty and item discrimination, it also accounts for guessing on test items. The one- and two-parameter models do not account for the possibility that individuals

are answering test items without error related to distractions or guessing. The item characteristic curve varies from the two-parameter model in that it allows the curve to begin somewhere greater than zero probability of answering the item correctly and vary based on test item. Three-parameter models are most frequently utilized in the education field.[144]

Partial Credit and Graded Response Models

The one-, two-, and three-parameter IRT models assume binomial response options for test items. More recent models have been developed which support polytomous response options, including the partial credit model and the graded-response model. Although these models can accommodate for alternate response options, they rely on the same parameters as the one- and two-parameter models.[163] The graded-response model is an extension of the two-parameter IRT model, it accounts for multiple response options by generating a characteristic curve for each between response option category threshold. This threshold represents the ability level needed to respond beyond a threshold of 0.50. The partial credit model is an extension of the one-parameter and Rasch model for polytomous response options. In this model, intersection parameters are utilized to represent the ability level at which an individual is more likely to respond to one response option category than another.[157] The partial credit model assumes equal distance between response options, while the graded-response model does not.[144,157]

Rasch Analysis

Overview of the Model

The one-parameter logistic model is sometimes referred to as the Rasch model. However, Boone and colleagues suggest that the Rasch model should not be considered an IRT model because of the many differences between the two. The fundamental difference, according to the

authors, is that IRT models add parameters to fit the data whereas in Rasch measurement the model is not altered to fit the data.[165] Using Rasch analysis, an outcome measure is tested against the mathematical measurement model to determine the fit of the data to the model, whereas in other IRT models the model is fit to the data.[166] Despite the apparent similarities between the one-parameter IRT model and the Rasch model, it is incorrect to assume they are the same.[165]

Assumptions specific to the one-parameter IRT and "Rasch model" include all items have equal discriminating power and the responder does not guess when answering test items. It is known that these assumptions are unrealistic, and Hambleton and Swaminathan suggest that the model may be robust to violations in these assumptions.[164] The Rasch model requires a smaller sample size (50-200 subjects depending on specific needs) than the two and three-parameter models, making it easier to use in clinical resarch.[142,167]

Selection of Model for the Study

This research study will utilize the partial credit Rasch analysis model. Rasch analysis is the simplest of the three models, but is also the most frequently used in health care. It is quickly gaining popularity in physical therapy research.[6-9,12,168] Rasch analysis allows for estimation of ability level based on item difficulty. While it does not account for item discrimination, observation of Rasch output does allow one to gain an understanding of the test's ability to separate individuals based on ability level through other methods described later. In addition, the outcome measures utilized in physical therapy-related research, and specific to this study, are based upon an individual's personal experience and should therefore not be arrived at through guessing. These factors rule out the need for the two and three-parameter IRT models. The partial credit model is necessary because the measures utilized in this study utilize polytomous response options.

Data Provided by the Model

There are several software packages that perform Rasch analysis, and the output of each varies slightly across programs. Winsteps (Winsteps® Rasch measurement computer program, Beaverton, USA),[169] RUMM2030 (RUMM Laboratory, Australia),[170] and ConQuest (Australian Council for Education Research, Camberwell, Australia)[171] are three of the most frequently utilized.[166] Output from Winsteps will be used in this study and described here. This section will provide a detailed summary of the data provided by Winsteps and Rasch analysis, including information related to scale dimensionality, person and item fit, differential item functioning (DIF), response scale structure and person and item reliability and separation indices.

Scale Dimensionality

As mentioned previously, an assumption of IRT and Rasch analysis is unidimensionality. The presence of unidimensionality allows for greater confidence that the scale is measuring what it is intended to measure (construct validity). Principle component analysis (PCA), person and item fit, DIF each contribute to the understanding of a scale's dimensionality, and therefore construct validity.

PCA in Rasch analysis looks for unexpected patterns in a group of test items. The presence of a group of test items that behaves unexpectedly when compared to the model could indicate the presence of an additional construct in the PRO.[172] The data are deemed to fit the model well when greater than 50% of the observed raw score variance is accounted for by the model.[173] An unexplained variance by the first factor eigenvalue of 1.4 to 3.0 or less suggests random correlation between remaining variables.[6,14,173] An eigenvalue is similar in concept to explained variance. A large eigenvalue, defined as a value greater than one, represents a strong source of variance or variability in a set of observed variables. Raykov and Marcoulides suggest

that the number of factors in a measure corresponds with the number of eigenvalues higher than one.[174] It is likely that internal and external factors, such as cultural difference, test anxiety or reading comprehension, introduce additional constructs to the measure, therefore changing the dimensionality. In general, there must be a dominant factor related to the ability measured within the test.[146]

In the case that PCA suggests the presence of more than one dimension in a measure, individual item loading coefficients (≥ 0.40) are considered. Groupings of test items in the positive or negative direction may suggest the presence of additional constructs in the measure, which could be analyzed as separate subscales in further PCA.[172]

Person and Item Fit

In Rasch analysis, logits are used to express both person and item measures. Person measure refers to how an individual performs on a specific test item and is interpreted as ability level. Item measure refers to item difficulty.[175] Wright Maps, also referred to as person-item maps, are used to display person and item measures, expressed in logits, on a single linear scale for a single variable using a scale hierarchy for test items. A Wright Map allows a researcher to visualize how the latent construct is defined through the range of ability levels addressed by the test. It also provides information regarding floor and ceiling effects, and the presence of redundant test items addressing the same ability level or the lack of test items (gaps) to capture an ability level.[176] Gaps in the item hierarchy suggest imprecision of the measure.[9] Valid measurement requires consistency in test item difficulty or person ability level despite who is responding to the item or which test items are being answered.[156] A variation of the Wright Map, the Rasch half-point threshold map, is available, which provides a more accurate representation of the polytomous representation of test items across ability levels.

The term "fit" in Rasch analysis refers to goodness of fit, or how well a data set conforms to the Rasch model. The Rasch model can predict, based on item difficulty level and person ability level, how a responder will answer a specific test item. Deviation from this predictable model would suggest misfit. Misfit on the person level occurs when an individual deviates from an expected pattern of test item response based on their ability level.[177] Person misfit could be a result of "aberrant response patterns" such as unrecorded comorbidity, including cognitive deficits.[166] For example, misfit occurs when responders with high ability level incorrectly answer a test item with low difficulty level, or vice versa. Misfit of persons and/or items degrades the quality of measurement. Person and item fit are further defined by infit and outfit statistics. Outfit statistics are more sensitive to outliers and may be a result of coding error. Infit statistics are sensitive to violations close to the ability level and pose a challenge to interpretation, and therefore a greater threat to validity.[177] According to Chiu and colleagues, outfit statistics are unweighted and occur far from the person or item, for example a person of low ability unexpectedly answers a difficult item correctly. Infit statistics are weighted and are affected by incorrect responses close to the person, item, or measure. For example, a person with low ability answers an easy or less difficult item incorrectly suggesting even lower ability.[7]

The presence of person or item misfit is determined by analysis of Person Infit MNSQ (mean-square) and ZSTD (z-standardized), Person Outfit MNSQ and ZSTD, Item Infit MNSQ and ZSTD, and Item Outfit MNSQ and ZSTD. MNSQ is a chi-square calculation in which the distribution of observed frequencies is compared to the expected frequencies of the Rasch model. If the observed and expected frequencies are the same, the chi-square statistic will be equal to zero.[178] A MNSQ value of 1.0 logit is expected.[177] Values greater than 1.0 suggest underfit and suggests unexpectedly high variability (variance or noise) in the response pattern, while values

lower than 1.0 suggest overfit. Overfit indicates that the pattern that is too predictable or redundant, and an item that fails to differentiate individuals.[15,177] Most surveys or performance scales have inherent error, therefore when considering person and item fit in Rasch analysis, a range of 0.5 to 1.5 logits is defined as a reasonable fit of the data to the model.[177] When the MNSQ values fall outside of this range, the ZSTD statistic is analyzed. The ZSTD provides a t-test statistic that measures the probability that the MNSQ calculation occurred by chance. A ZSTD of > ±2 would suggest person or item misfit.[177] In scale development or revision, one may choose to eliminate misfitting items from a scale, however this could result in a change in the validity of the measure and interpretability and should therefore be done with caution.[179]

Differential Item Functioning

DIF refers to the "*relative difficulty* of individual test items for groups with dissimilar cultural or experimental backgrounds," but equal ability levels, that results in unequal probability of success.[158(p 196-7)] In many cases, authors attribute the presence of DIF to item bias.[12,166] According to Boone, however, DIF does not seek to identify whether test items are unfair or biased between groups, rather whether test items "operate in the same way for different groups of respondents."[180(p 274)] Different groups within a sample may answer test items differently based on their gender, age, race or language, for example, and if present could suggest the presence of additional constructs or dimensions tested within the scale.[180] DIF is assessed by comparing the item characteristic curve for the same item across two different groups.[158] A measurement instrument with high construct validity should not have test items that shift in order or spacing along the item hierarchy based on the group of people answering. Individual test items which are less than a probability of 0.05 and have an effect size of greater than 0.64 must be qualitatively considered for DIF.[180] In the event that DIF is present, resolution requires consideration of the

following: (1) ignoring it as inherent in the measuring system; (2) remove or rewrite the item; (3) treat data for one DIF group as missing data for further analysis; or (4) split the item making two test items, one item with active data for one DIF group and missing data for the other.[181]

Response Scale Structure

The structural integrity and scale functionality of a rating scale can be assessed in Rasch analysis by looking at response category utilization, uniform distribution of use of response options, scale calibration, and absence of misordered category steps.[8] Specifically, Linacre proposes five necessary characteristics of an optimal rating scale: (1) test items are oriented with the latent variable and are of the same scoring structure; (2) there are a minimum of 10 responses per response option for each test item; (3) there is uniform distribution of category utilization; (4) average measures per category and step calibrations increase monotonically; and (5) the response category outfit MNSQ does not exceed 2.0.[182] At least 10 observations per response category are needed to determine measure stability. Fewer observations per response option can result in imprecise step calibration and make the scale unstable.[182,183] Step calibration provides information related to ordered response options. Step calibration should sequentially increase or decrease with item difficulty. The average measure values should also increase or decrease with item difficulty for each response option. Failure of average measures to increase up the rating scale limits the ability to interpret the measure for the data set. Mean-square statistics calculated for each response category provides an estimation of the degree of error within a category An outfit value of greater than 2.0 suggests that there is more unexplained variance than explained variance in a category, and therefore limits the accuracy, stability and interpretability of the measure.[182] Scale structure can also be assessed through consideration the Rasch half-point

threshold map and the presence of gaps or redundancies in the rating scale and the presence of floor or ceiling effects.

General Scale Indices

Rasch analysis provides evidence for test reliability through assessment of dimensionality and the provision of an unbiased reliability estimate. Rasch analysis through Winsteps provides output for person and item reliability, and person and item separation, both of which are reported in logits. A logit is a "log-odds unit" derived from the log odds ratio of the probability of success divided by the probability of failure on a test item. Logits may range in value from zero to infinity, and are interval level data that have been transformed from ordinal level data.[184] Person reliability is related to the ability of the test to discriminate into ability levels. Person reliability in Rasch analysis is similar to internal consistency (Cronbach's alpha, α) in CTT where a value closer to one indicates greater internal consistency.[185] Smith[151] suggests differences between these values, however. In CTT, internal consistency is developed from data that is non-linear and includes extreme scores (perfect and zero scores). Inclusion of extreme scores (perfect and zero scores) in reliability calculation decreases the average error and inflates reliability reporting. Therefore, in Rasch analysis, extreme scores are often eliminated from analysis.[151] Elimination of extreme scores is necessary because the error of the person measure is "infinite." There is no way of knowing how far beyond the item score a person's ability level continues. Inclusion of these individuals does not provide any useful information about how accurately the test is functioning. [185] In Winsteps, the 'REAL' estimate provides the person reliability coefficient with extreme scores removed. The 'MODEL' estimate does not remove extreme scores. Boone and colleagues suggest using the REAL estimate when reporting.[185] A minimum value of 0.7 is preferred for group level analysis, and a value of 0.85 is suggested for the individual level.[166]

Item reliability is also reported from zero to one. Higher item reliability suggests the presence of a large range of item difficulty within the measure. Therefore, low item reliability may suggest an inadequate sample size to establish item difficulty hierarchy. Item reliability is independent of test length and uninfluenced by model fit. A reliability index of greater than 0.9 is desired to verify item hierarchy.[186]

Rasch output also provides a person and an item separation index.[185] The separation index is the square root value of the ratio between the true person variance and the error variance in the data.[173,185] The separation index can range from zero to infinity, with a higher score being more desirable. Boone and colleagues reference a discussion with Mike Linacre in 2012 in which Linacre explained that person separation is used to classify people into ability levels. A low person separation index may suggest that the measure is not sensitive enough to distinguish between ability levels. The item separation index verifies item hierarchy.[185] A person separation index of 1.50 is acceptable, a value of 2.0 indicates a good level of separation, and a value of 3.0 is excellent. For item separation, a value of 1.5 is required for analysis at the individual level and a value of 2.5 is required for group level analysis.[185]

Reliability and Validity – Establishing the Theoretical Framework for this Study

The increasing popularity of IRT and Rasch analysis offer additional methodologies to study the functionality and appropriateness of test score interpretation of a PRO. In general, CTT is "test based," whereas IRT and Rasch models are "item based." Item based models allow for a broader range of interpretations based on individual test performance.[145] Pusic and colleagues therefore suggest that CTT be used for group-based research, while IRT and Rasch analysis be used for interpretation on the individual level.[5] According to Tennant and Conaghan, Rasch analysis should be used in the following circumstances: when a set of ordinal-level test item

scores are intended to be combined into a composite score; in the development of a new scale; when reviewing the psychometric properties of existing scales; in the assessment of the dimensionality of tests; when constructing test item banks for computerized adaptive testing (CAT); and when change scores need to be calculated from ordinal scales.[166]

IRT and Rasch analysis are useful in supporting the reliability and validity of a PRO.[151] The section will provide an overview of the information gleaned from Rasch analysis regarding reliability. It will also introduce the theoretical framework for this research study related to construct validity and the ability of Rasch analysis to contribute to the validity of a measure's test score interpretation.

In Rasch analysis, the standard error estimate of a person's ability level and item difficulty level provide information related to reliability on an item level. In addition, the person reliability estimate provides information related to internal consistency and is analogous to Cronbach's alpha in CTT.[151] Item reliability is unique to IRT and Rasch analysis in that it offers a reliability estimate for the items[185] The reliability of a measure contributes to the validity of a measure. In other words, a measure must be reliable to be valid, however the measure does not need to be reliable to be considered valid.[187]

Although there are many forms of validity reported in psychometric models,[187] Messick suggests that validity is a unified concept in which the multiple facets of validity (content, substantive, structural, generalizability, external, and consequential) each contribute to the overall construct validity of the measure.[188,189] Construct validity is defined as "any evidence that bears on the interpretation or meaning of the test scores."[189(p 7)] The accuracy or appropriateness of test score interpretation is never definitively defined, rather the available evidence to support construct validity is considered to establish the "degree" to which a test score interpretation can

be trusted.[189] Various approaches must therefore be utilized to contribute to the evidence for validity.[151]

Rasch analysis has three components that address the six facets of construct validity proposed by Messick (and listed previously)[188,189] and two additional facets (responsiveness and interpretability) proposed by the Medical Outcomes Trust[151]: (1) model requirements and measurement properties if the data fit the model; (2) linear hierarchical scale and standard error; and (3) item and person fit.[151] The eight facets of validity and how they are assessed using Rasch analysis are as follows[151,188,189]:

- The content facet of reliability refers to the relevance, representativeness, and technical quality of a measure. The relevance and representativeness are substantiated by analysis of the item hierarchy and item calibration. Failure of a scale to cover a full range of ability levels related to a construct suggests poor interpretability and utility. Technical quality is addressed through analysis of fit statistics, and the degree that each test item contributes to measuring the latent variable.
- The substantive component is confirmed through analysis of the item hierarchy and whether the hypothesized and empirical item difficulties are similar. Person fit also contributes to this facet.
- The structural facet refers to the credibility and interpretability of the scoring system. It is supported by item and person parameters.
- The generalizability aspect of construct validity addresses the invariance of person measures and item calibration across different groups, times or contexts. It can be addressed through comparison of means for different samples. Generalizability will not be assessed in this research study.

- The external facet is concerned with convergent and discriminant evidence. Discriminative evidence can be addressed through assessment of variance between known groups, such as with DIF. Convergent evidence can be assessed through calibration and comparison of person measures with another measure.
- Consequential evidence considers the value and consequences of score interpretation and test use. It also considers test fairness, which can be assessed using DIF.
- Responsiveness is related to the measure's ability to detect change. In Rasch analysis scale responsiveness can be addressed by assessing the ability of the measure to distinguish between two ability levels using person reliability and separation index.
- Interpretability refers to the ability of a quantitative meaning to be assigned to a qualitative measure. In Rasch analysis, this is determined through analysis of the person-item map because it places item difficulty and person ability on the same scale. This allows for prediction of how an individual will answer an item of specified difficulty level based upon their ability level.

According to Messick, validity refers to the "relevance and utility of the test scores for particular applied purposes and of the social consequences of their use."[189(p 5)] Two sources of construct invalidity include construct underrepresentation and construct-irrelevant variance. Construct underrepresentation occurs when the measure fails to include all relevant aspects of a construct; and construct-irrelevant variance occurs when the measure is too broad and includes variance related to other dimensions or constructs.[188,189]

Patient-Reported Outcome Measures to Objectify HNC-Related Shoulder Dysfunction

Introduction to the Section

As stated previously, physical therapists should utilize validated PROs to quantify shoulder-related impairment and to demonstrate value of interventions provided. There are numerous shoulder-related PROs available to the clinician, only a few of which have been utilized in HNC research. Even fewer were developed specifically for the HNC patient population. This section will summarize the systematic reviews available in the literature that have attempted to offer clinicians treating this population with guidance as to which PROs are most appropriate. An in-depth analysis of the literature related to seven shoulder-related PROs will then be provided.

Recommended PROs for HNC-related Shoulder Dysfunction

When considering shoulder dysfunction in the HNC population, there are very few PROs that have been developed for and validated in populations of individuals presenting with shoulder dysfunction following neck dissection surgery. In addition, the psychometric research that has been done is grounded in CTT, thus is susceptible to the limitations previously outlined. Researchers have attempted to provide medical professionals treating patients with HNC with recommendations of the most appropriate shoulder-related PROs for the population.[2,3,190] Through a review of HNC literature published between 1980 and 2011, Goldstein and colleagues identified seven PROs that have been utilized in research to quantify HNC-related shoulder dysfunction, including: modified American Shoulder and Elbow Surgeons standardized form (ASES), CS, Disability of the Arm, Shoulder and Hand (DASH), NDII, Shoulder Disability Questionnaire (SDQ), SPADI, and the Simple Shoulder Test (SST). Of these measures, the NDII is the only measure identified that was specifically designed and validated in the HNC

population, however the measure's psychometric properties have not been thoroughly researched. The authors highlight the DASH as the most extensively developed measure across diagnoses, however also point out that the psychometric properties of the measure have not been established in the HNC population.[3]

The Previously Untreated, Locally Advanced (PULA) Task Force of the Head and Neck Steering Committee of the Coordinating Centre for Clinical Trials at the National Cancer Institute has provided recommendations for PROs use in clinical trials for HNC.[190] Similar to the Goldstein article,[3] the Task Force also recommended the DASH and NDII as subjective measures of shoulder and tissue fibrosis. The authors also named the CS as a measure of shoulder and tissue fibrosis, however do not include it as one of the recommended measures. Their rationale for this is not stated.[190]

The Academy of Oncologic Physical Therapy of the American Physical Therapy Association's (APTA) Evaluation Database to Guide Effectiveness (EDGE) Task Force[191] for HNC has also issued recommendations for physical therapists treating patients with HNC-related shoulder impairment based upon an extensive literature review of outcome measures utilized across patient populations and diagnoses to quantify shoulder impairment. Nearly 50 PROs related to the shoulder were found, and 16 met the inclusion criteria for analysis by the Task Force. The Task Force recommendations are based upon a review of the psychometric properties, based on CTT, of the tools, the population and condition the tool was developed in, and whether the tool has been used in HNC-related research. Table 2 provides a description of the recommendation system. The Task Force highly recommends (score of 4) the DASH, the SPADI, NDII, and the UW-QoL (shoulder subscale) for use by physical therapists when quantifying shoulder dysfunction following neck dissection surgery, and recommends (score of

3) the QuickDASH. The Task Force chose to recommend the UW-QoL (shoulder subscale) and the NDII because they were developed specifically for the HNC population and have shown promising reliability and validity. The DASH, QuickDASH, and the SPADI, on the other hand, were not developed in the HNC population, but have been utilized and tested extensively across multiple patient populations and diagnoses.[2] The DASH and the SPADI have both been used in HNC research,[21,66,79,86,109,114,115,128,192-199] but have not been sufficiently tested to demonstrate adequate psychometric properties.

Table 1. Head and Neck Cancer Rating Scale

4	Highly Recommended	Highly recommended; the outcome has good psychometric properties and good clinical utility; the measure has been used in research on individuals with or post head and neck cancer.
3	Recommended	Recommended; the outcome measure has good psychometric properties and good clinical utility; no published evidence that the measure has been applied to research on individuals with or post head and neck cancer.
2A	Unable to Recommend at This Time	Unable to recommend at this time; there is insufficient information to support a recommendation of this outcome measure; the measure has been applied to research on individuals with or post head and neck cancer.
2B	Unable to Recommend at This Time	Unable to recommend at this time; there is insufficient information to support a recommendation of this outcome measure; no published evidence that the measure has been applied to research on individuals with or post head and neck cancer.
1	Do not Recommend	Poor psychometric and/or poor clinical utility (time, equipment, cost, etc.)

A Review of the Psychometric Properties of Shoulder-Related PROs

American Shoulder and Elbow Surgeons Standardized Form

The ASES, first published in 1994, was developed by a research committee of the American Shoulder and Elbow Surgeons in an effort to facilitate communication between researchers and encourage multicenter trials. The test was developed through assessment of all published outcome measures at that time and expert opinion from the members of the organization.[200,201] The ASES has 2 sections – a clinician assessment (cASES) of ROM, strength

and instability, which is not included in the total test score, and a patient self-evaluation (pASES). In most cases, the pASES is the only section reported in research. The pASES consists of 11 items, which are divided into two areas, pain [one item scored on a 10-cm visual analog scale (VAS) with 10-cm indicating worst pain ever] and function (10 items). The function items are specific to ADLs, usual work and recreational activities, and are unique in that they ask the responder to rate the difficulty separately for the left and right upper extremity. The recall period for the pASES is one week. Test items are based on a 4-point Likert scale (0 = unable to do so, 3 = not difficult). The pain scale response is multiplied by five to get a score out of a maximum score of 50, and the function subscale is multiplied by 5/3, to also provide a score out of a maximum score of 50. An overall score of 0-100 points is possible, where a score of 100 indicates normal shoulder function. The pASES has low responder and tester burden, taking approximately 3-5 minutes to complete and two minutes to score.[149,202,203] Bot and colleagues have classified the ASES as difficult to score, however.[204] A modified version of the ASES added five additional questions related to pain, including the use of pain medication and shoulder instability. The newer test items are not included in the test score, however.[205] The ASES is available in German,[206] Italian,[207] Turkish,[202] Arabic,[208] Finnish,[209] Spanish,[210] and Portuguese.[211,212]

The ASES was developed to provide a measure of shoulder function irrespective of diagnosis.[201,203] It has also been utilized in studies addressing shoulder pain and dysfunction,[202,203,212] subacromial impingement,[210,213] rotator cuff dysfunction[210,214] instability,[207,214] and rheumatoid arthritis (RA), osteoarthritis (OA), total joint arthroplasty,[206,214-217] and clavicle fracture.[218] Normative values range from 92-99 points.[219,220]

Reported test re-test reliability and internal consistency in the literature ranges from Intraclass Correlation Coefficient (ICC) = 0.75-0.96,[202,203,206,207,209,212,214,217,219,221,222] and Cronbach's α = 0.61-0.96,[202,203,206,207,210,212,214,222] respectively. Person reliability, which is analogous to internal consistency in Rasch analysis, is reported as 0.89 – 0.90.[20,210] An SEM of 2.51,[202] 10.55,[212] 11.0,[203] is reported. The Minimal Detectable Change$_{95\%}$ (MDC) can therefore be calculated as 6.96,[202] 29.24,[212] and 18.74.[203] Minimal Clinical Important Difference (MCID) values range from 6.4-16.9 in the literature.[203,223] The Turkish and Spanish versions were found to have minimal to no floor and ceiling effects.[202,210] However, significant ceiling effects are present in the pain and instability sections of the measure for patients who have received total shoulder arthroplasty.[216]

Convergent validity is reported using the Spearman's correlation coefficient (r) for parametric data or the Pearson's correlation coefficient (r_s) for nonparametric data. For simplicity both Spearman's and Pearson's correlation coefficient will be presented as 'r' in this document. The convergent validity of the ASES has been established with comparison with other shoulder-related outcome measures as follows: SPADI total score (r = -0.81 to -0.92),[202,216] DASH (r = -0.47 to -0.92),[206,207,212,216,217,224] QuickDASH (r = -0.55 to -0.85),[224] SST (r = 0.536-0.73),[209,225] the physical functioning, role limitation, and bodily pain domains of the Medical Outcomes Short Form 36 (SF-36, r = 0.33-0.74),[203,209,210,212,216] CS (r = 0.71-0.87),[216,226] Turkish version of the CS (r = 0.48),[202] Spanish version of the CS (r = 0.62)[210] Barthel Index (r = 0.31),[210] and the Penn Shoulder Score (r = 0.78-0.87).[203,227] Discriminant and divergent validity have also been supported.[203] Responsiveness of the ASES has been reported with a standardized response mean (SRM) of 1.6 and an effect size (ES) of 1.4.[203] Partial credit Rasch analysis of a modified version of the pASES performed in 2001 found misfit of two test items, Item 1: *sleep on shoulder* and

Item 10: *reach behind back to fasten brassiere*. The analysis suggests that the pASES does not have equal interval measures, indicting a limitation in responsiveness to change over time. In addition, the measure is imprecise at measuring those with lower and higher shoulder functioning.[20]

Rasch analysis was utilized to validate the Spanish version of the ASES. Due to the concern of multidimensionality, the authors utilized factor analysis and Rasch analysis on two models: (1) all 11 items and (2) the 10 function-based items (eliminating the pain analog scale). Analysis supported unidimensionality of the full 11-item scale. Person and item reliability (0.90 and 0.98) and person and item separation indices (2.93 and 7.53) were deemed acceptable. Item 10: *do usual sports* had the highest frequency of missing data and poorest fit to the model. Item 7: *lift 10 pounds over shoulder* and Item 8: *throw a ball overhand* were the most difficult to endorse and Item 4: *manage toileting* was the easiest to endorse.[210] Beckmann and colleagues utilized Rasch analysis to study the 11-item scale and report limitation related to multidimensionality and poor person reliability (0.48). Strengths included excellent item reliability (0.98) and no floor or ceiling effects.[225]

Wright and Baumgarten suggest that the ASES is the most appropriate measure to be used for research purposes as a general shoulder measure.[228] Schmidt and colleagues performed a comprehensive review of PROs addressing shoulder dysfunction in a broad range of diagnoses, and also supported the ASES as the best overall shoulder-specific PRO.[229] The ASES has been used in HNC literature,[116] however psychometric properties specific to the population are not reported.

Constant's Shoulder Score

The CS, also known as the Constant's Score and the Constant-Murley Score, is one of the original shoulder-specific PROs first published in 1986 and 1987. It was proposed by orthopedic surgeons in the European Society for Surgery of the Shoulder and Elbow as a shoulder-specific measure to compare shoulder function before and after treatment,[228] and according to the developers is "applicable irrespective of the details of the diagnostic or radiologic abnormalities caused by disease or injury."[230(p 160)] The methodology for development of the CS, including item selection, item reduction and item weighting, has not been published.[200] The measure was further reviewed and modified in 2008,[231] and a standardized testing protocol was published in English and Danish in 2013.[232]

The CS consists of a patient-reported section, which accounts for 35% of the total test score, and a clinician measure that accounts for the remaining 65% of the test score. The patient-reported section includes five items that address pain and ADLs. The first item addresses the most severe pain experienced during ordinary activities within the past 24-hour period, using a VAS with anchors of 'no pain' and 'intolerable pain.' Pain is scored on a 0-15 point scale, where a pain level of zero is awarded 15 points. The ADL section has 20 available points distributed over four test items related to sleep (0-2 points, 2 points awarded for undisturbed sleep), work disturbance (0-4 points, 4 points awarded for no limitation), recreational activities (0-4 points, 4 points awarded for no limitation), and the ability to lift the arm (0-10 points, 10 points awarded for lifting above the head, 8 points to head, 6 points to neck, 4 points to sternum, 2 points to waist, and no points for below the waist). The remaining 65 points are scored based on the clinician's assessment of shoulder ROM (40 points) into forward flexion, abduction, internal and external rotation measured with a goniometer and strength (25 points) measured at

90° of abduction in the scapular plane using an Isobex isometric dynamometer or myometer. The average of three trials performed with one-minute rest between trials is scored. A subject who cannot achieve 90° of abduction receives a score of zero. A score of 100 indicates no limitation.[231] Individuals are asked to recall their symptoms over the past week when completing this questionnaire.[205] The CS takes 10-15 minutes to complete and two minutes to score.[149,228] According to Angst and colleagues, the CS is used in most languages without official translation,[205] but has been validated in the Turkish language.[202] Performance of the test on both extremities allows for comparison between sides.[233]

The psychometric properties of the CS have been reported in various patient populations with musculoskeletal disorders and shoulder instability. Normal CS values for male and female subjects, aged 21-100 years old, are reported in the literature.[231] The test-retest reliability is ICC = 0.80-0.96.[3,149,202] Internal consistency (α) ranges from 0.37-0.60.[205] The strength domain provides a high risk of error in measurement resulting in low intertester reliability.[205] MDC, MCID, and SEM are not reported.[3,205] Effect Size (ES) and SRM for patients with OA and RA undergoing shoulder arthroplasty, shoulder impingement with surgical decompression, rotator cuff repair, and physical therapy for instability are reported.[205]

The convergent validity of the CS is supported through the correlation (r) with several PROs including, the ASES (0.48-0.87), Oxford Shoulder Score (0.65-0.87), DASH (-0.50 to -0.82), SPADI (-0.53 to -0.82), and the SST (0.49).[202,205,216,234] A review of PROs cites no floor or ceiling effects for the CS total score,[202,205] however, significant ceiling effects were found for the pain and ADL components of the CS in a population of patients receiving total shoulder arthroplasty.[216] A floor effect as high as 52% has been reported for the strength subscale because subjects were unable to achieve the required 90° abduction, and therefore were awarded a score

of zero.[235] This is a significant finding and a consideration for the HNC population given the tendency for abduction ROM to be significantly impaired. Content and divergent validity however are described.[3]

The CS has been used in research related to SAN injury,[236] and in HNC-related studies.[65,74,116,123,129,130,237-241] According to Goldstein,[3] normative data are not available, however, the validation study of the NDII reports a mean (SD, range) score in a population of patients with HNC as 70.7 (17.4, 38-100).[4] Similar findings are reported by Chepeha and colleagues in a population of patients 11 months following SND and MRND surgery for HNC (71.0 ± 18.8, 22-100).[238] Watkins and colleagues utilized cut off scores to classify subjects as having mild, moderate, and severe shoulder dysfunction by comparing the involved extremity to the uninvolved extremity. Subjects with 1-2 SD variation between sides were classified as having mild dysfunction, and subjects with 2-3 SD and greater than 3 SD were classified as moderate and severe dysfunction respectively.[65] Convergent validity in HNC populations has been established with correlation of the CS with the NDII ($r = 0.85$[4] and 0.64[129]). The subjective and objective measurements of the CS are also significantly correlated ($r = 0.65$) in a population of patients with HNC.[238] Merve and colleagues support the discriminative validity of the CS based on neck dissection type.[74] Psychometric properties related to reproducibility and responsiveness to change has not been reported for the HNC population.

A limitation of the CS is the required equipment for testing strength, which may not be readily available in the clinic. According to Constant and colleagues, manual muscle testing is "condemned."[231] Other limitations include variations in CS administrative practices, limited evidence to support reliability and validity across patient populations, lack of normative values for various patient populations, and lack of MDC and MCID values.[205] Several authors suggest

the CS scoring system may be inappropriate due to the multidimensionality of the clinician- and patient-rated domains, and should therefore not be included in the same test score. There is also criticism based on the weighting of the various items within the scale. ROM accounts for 40% of the scale, while strength accounts for 25%, pain 15% and function 20%. In addition, the risk of bias based on variations in the clinician assessment has also been suggested as a weakness of the scale.[200,228] It does not appear that the psychometric properties of the CS have been assessed using Rasch methodologies at this time.

Disabilities of the Arm, Shoulder and Hand, QuickDASH

DASH

The DASH was originally developed in 1996 by the American Academy of Orthopedic Surgeons and the Institute for Work and Health as a "tool to be used for patients with any condition of any joint of the upper extremity."[200(p 1113)] Test items were generated through a review of the literature to establish an item pool of 821 items that were reduced through expert opinion, pilot testing in a sample of 20 patients, and field-testing in a sample of over 400 subjects across various upper extremity diagnoses and regions.[3,200,242] Information about the DASH, including test development, populations tested, and psychometric properties, is available in published manuals. The most recent was published in 2011.[243] The DASH has been translated into 52 versions, with 20 additional versions currently in progress.[244] Two shortened versions of the DASH, the QuickDASH and the QuickDASH-9, are also available.[245,246] The QuickDASH is widely used in clinic and research settings, whereas the QuickDASH-9 is rarely used and is not supported by the developers of the DASH and QuickDASH.[205] MacDermid and colleagues suggest the QuickDASH adequately covers the range of difficulty of DASH items and therefore may be an appropriate "surrogate" for the DASH.[224] In 2017, Kennedy and Beaton published the

results of a survey of DASH and QuickDASH users. The purpose of the survey was to: (1) determine how the measures are being utilized in clinical and research settings, and (2) to understand which items users found most useful and most problematic to better inform further iterations or versions of the measure. The authors concluded that the DASH and QuickDASH are being utilized as intended: in more than 29 countries; in various practice settings; in patient care and research; across all diagnoses affecting the upper extremity; and populations of all ages.[247]

The DASH is a 30-item PRO designed to quantify disability related to the entire upper extremity.[244] It includes five items which address body function and structure, 21 items which address activity limitation, and four items which address participation restriction.[248] Franchignoni and colleagues suggest an alternative division of test items (constructs) as follows: manual functioning (items 1-5, 7-11, 16-18, 20, 21), disability because of limitation in shoulder ROM (items 6, 12-15, 19), and symptoms and effects of the patient's problem (items 22-30).[14] The DASH also includes two optional 4-item modules for work and sports/performing arts that are scored separately.[244] The recall period is one week. Each item is scored on a Likert scale of 1-5. A score of one represents 'no difficulty' and 'not limited at all' and a score of 5 represents responses of 'unable' or extreme difficulty. Scoring requires summing the 30-item responses, subtracting 30 from the total, and dividing by 1.2. Missing items are replaced by the mean value of the other responses before summing. The score cannot be calculated if more than three items are missing.[228] An alternative, more frequently used scoring system is also available [(raw score/number of items measured) – 1] x 25.[244] A total test score of 0-100% is possible, where a higher score indicates greater dysfunction. The DASH takes 5-7 minutes to complete and three minutes to score.[149,228]

Carr and colleagues published a classification system for the DASH in which a score of 1-33% is characterized as mild disability, a score of 34-67% is characterized as moderate disability, and a score of 68-100% is suggestive of severe disability.[66] This classification system is not widely utilized, nor is it acknowledged on the DASH website.[244] Kennedy and colleagues published a different set of cut-off scores, in which a score of less than 15 is interpreted as 'no problem,' a score of 16-40 represents 'problem, but working,' and a score of greater than 40 suggests 'unable to work.'[243] Normative values (population mean ± SD) from a sample of 1706 individuals in the United States have been reported as 10.1±14.68.[249] Aasheim and Finsen reported similar findings in Norway with a mean population DASH score of 13. The authors found increasing scores with age, and higher scores in women (15±3) than men (11±2). Age and gender based normative values are available in the literature.[111] Normative values in a population of young, active adults with a mean age 28.8 years has, however, been reported to be lower than the population mean (1.85±5.99).[220]

The original version of the DASH was developed in a population of 420 patients, age 18-65 years, presenting to 23 outpatient clinics across Canada, Australia and the United States with disorders of the upper extremity. Upper extremity diagnoses included Colles' fracture, carpal tunnel syndrome, symptomatic OA, RA, painful arc of the shoulder, lateral elbow pain, and nonspecific soft tissue pain.[242] The DASH has since been utilized in the literature for numerous populations, including: RA,[250-252] psoriatic arthritis and inflammatory disease,[253] post-operative upper extremity surgery,[250] intercollegiate athletes,[248] non-traumatic neck complaints (new onset or recurrence) with upper extremity symptoms,[254,255] multiple sclerosis (MS),[11] shoulder impingement,[256] subacromial pain syndrome,[257] adhesive capsulitis,[258] proximal humeral fractures,[259] "any upper extremity problem (excluding the shoulder),"[16] and stroke.[18] The DASH

has also been utilized in research related to traumatic brachial plexus nerve injuries,[260] and breast cancer.[261,262] The psychometric properties of the DASH have been found to have excellent reproducibility, with internal consistency/cross-sectional reliability ranging from Cronbach's α = 0.87-0.98[11,18,149,205,249,252,261] and test-retest reliability ranging from ICC = 0.91-0.99.[149,205,250-252,256,257,259,263,264] Responsiveness to change of the DASH has also been reported in various populations. In a general musculoskeletal population a SEM = 4.3-7.6 and $MDC_{90\%}$ = 10.81-12.8 have been reported.[149,256,257,264] A $MDC_{90\%}$ of 10.7 and a $MDC_{95\%}$ of 7.9 – 14.8 have also been reported.[217,256,263,265] An MCID of 10.83 points is reported in a population of patients with upper-limb musculoskeletal disorders.[264] Area under the receiver operating characteristic curve (ROC-curve) is reported as 0.67[266] and 0.77 (95% CI: 0.63, 0.92).[257] ES and SRM are reported in the literature for various musculoskeletal diagnoses.[205]

Convergent validity of the DASH is supported by multiple comparisons with other shoulder-related and generic HRQOL PROs: SPADI (r = 0.55 - 0.93),[216,217,258,267,268] pASES (r = -0.63 to - 0.79),[216,217] CS (r = -0.82),[216] VAS (0.31),[258] Upper Limb Functional Index (ULFI) (0.90),[266] QuickDASH (0.94-0.99),[224] and SF-36 physical function scale (r = 0.67).[216] Beaton and colleagues support the discriminative validity of the DASH in a sample of adults with upper extremity complaints. Participants who were working with their upper limb condition and were able to continue to work had a significantly lower disability than those unable to work (26.8 vs. 50.7, t=-7.51, p<0.0001). Similarly, the DASH was able to discriminate between those who could do everything they wanted to compared to those who were not (23.6 vs. 47.1, t=-5.81, p<0.0001).[263] Low to no floor or ceiling effects are reported in populations of patients with RA, proximal humeral fracture, total shoulder arthroplasty, adults with subacromial pain, and adults with musculoskeletal upper extremity complaints,[216,252,257,259,263] however, a significant ceiling

effect was found in a population of intercollegiate athletes in which 65% scored a zero ('no disability').[248] In a population of individuals with upper extremity impairment (excluding the shoulder) a 4% ceiling effect (score of 0) and a 1% (score of 100%) floor effect is reported.[16] Through Rasch analysis, discussed in more depth later, researchers have questioned the construct validity of the DASH.[11,14,205]

Kennedy and Beaton surveyed 157 DASH and QuickDASH users. Users selected the following five items as most informative: Item 1: *open a tight jar*; Item 16: *use a knife to cut food*; Item 23: *limited work or daily activities due to upper extremity problem*; Item 29: *difficulty sleeping*; and Item 6: *place and object on a shelf above your head*. The following five items were selected as most problematic: Item 21: *sexual activities* (most problematic); Item 18: *recreational activities with some force*; Item 20: *manage transportation needs*; Item 1: *open a tight jar*; and Item 7: *do heavy household chores.* Of note, Items 21, 18 and 20 have been flagged by the authors as items needing further consideration in future iterations of the DASH.[247] Item 21*: sexual activities* on the DASH has previously been reported in the literature as being a problematic item.[111,205]

MacDermid and colleagues provide QuickDASH and DASH item ranking by difficulty for individuals prior to surgery and three and six months following total shoulder arthroplasty or rotator cuff repair. For both surgical procedures Item 19: *recreational activities in which you move your arm freely* and Item 18: *recreational activities in which you take some impact* are deemed the most difficult whereas Items 2: *write* and Item 3: *turn a key* are the easiest.[224]

The DASH has been used to assess shoulder disability and function in individuals with HNC,[66,86,192,193,196-198] and has been used in studies to assess shoulder function with other diagnoses of cancer affecting the upper extremity,[193,261,262,269] including individuals with thyroid

cancer following ND surgery.[270] It has also been used to study upper extremity disability in individuals undergoing radial forearm flap harvest for HNC.[271] Reliability and validity of the DASH in a population of patients undergoing unilateral neck dissection found a test-retest reliability of ICC = 0.91 (95% CI, 0.90-0.98) and convergent validity based on a strong relationship with the NDII of -0.86.[196]

The DASH was used to quantify shoulder dysfunction in a population of 65 patients (mean age 62 years, 77% male) 1.6 years (range, 0.5-4 years) following various SND surgeries. The authors provide normative values for the study population of mean DASH score of 21.1 ± 23.3 (range, 0-97.5). Despite the variability in the sample, the authors suggest that the DASH has high sensitivity in this patient population and recommend that the DASH be utilized to establish baseline shoulder function for post-operative comparisons and rehabilitation planning. The authors however do not provide psychometric data related to reliability, validity or responsiveness to change in this study.[66] Another study utilized the DASH to quantify shoulder dysfunction in a population of individuals with nasopharyngeal cancers after undergoing SND procedures, including dissection of level IIB. In this group, individuals scored mean 4.2 (SD, 1.8) prior to surgery, 44.2 (10.1; range, 28.0-66.5) one year after surgery and 46.4 (12.4; range 22.3-70.5) two years after surgery. The authors report the most frequently affected movements in this population (test items answered as being most difficult) as Item 6: *place an object on a shelf over the head*, Item 13: *wash and blow dry your hair,* and Item 15: *put on a pullover sweater*.[86]

QuickDASH

The QuickDASH, developed in 2005, is a modified version of the DASH developed to decrease responder and tester burden. The QuickDASH was developed using three modern questionnaire development strategies for item reduction: concept-retention, equidiscriminative

item-total correlation approach, and Rasch analysis.[205,245] The three questionnaire variations derived from these strategies were compared and ultimately the concept-retention version was maintained as the new QuickDASH. Based on the concept-retention strategy, DASH items were ranked according to the importance and difficulty of each item based on field-testing, and the item correlation with the overall test score. Domains related to weakness, stiffness, family care, sexual activities, and self-image were ellimiated.[245]

Like the DASH, the QuickDASH is a region-specific PRO. It consists of 11 items taken directly from the DASH (two related to symptoms, eight related to function) and the two optional modules also included in the DASH. The recall period is one week. Ten of the 11 items must be completed to score. The QuickDASH is scored using the same methodology as the DASH, yielding a final score of 0-100% disability related to the upper extremity.[244] It has been translated into 48 different languages,[244] and is gaining popularity in the musculoskeletal patient populations.[245,246,255,272-283] It also been validated in a population of breast cancer survivors.[284]

Population based (Norway) normative data based on age and gender have been published, and are similar to the DASH data.[111] Psychometric properties of the QuickDASH, including reliability and validity, have also been shown to be similar to the DASH.[245,285] QuickDASH internal consistency ranges from $\alpha = 0.91$ to 0.95 and test-retest reliability (ICC) ranges from 0.90 to 0.94.[205,264,284,286] Test-retest reliability in the breast cancer population is somewhat lower, with a reported ICC of 0.78.[284] Convergent validity of the QuickDASH is supported through high correlations with other shoulder specific PROs, including the SPADI (0.84),[205] DASH (0.93-0.99),[111,224,245,287] ASES (-0.55- -0.85),[224] ULFI (0.86),[272] VAS for function (0.80), VAS for pain (0.64-0.73),[286] and global HRQOL measures including the SF-36 physical functioning scale (0.68).[205] Ability to detect change for the QuickDASH is reported as $MDC_{90\%} = 11.0\text{-}17.2$[264,286]

and $MDC_{95\%}$ = 12.63-24.7 based on population.[276,284,286] The MCID in a population of patients with upper-limb musculoskeletal disorders is 15.91 points.[264] ES and SRM for musculoskeletal populations are reported,[205] and indicate that the QuickDASH is sensitive to varying amounts of change. However, poor correlations with estimates of change suggest low responsiveness of the measure.[286] Chester and colleagues suggest that the QuickDASH is able to discriminate between physical therapy responders and non-responders in a population of individuals with shoulder pain, reporting an area under the curve (AUC, 95% CI) of 0.78 (0.75, 0.81) at 6 weeks following initiation of physical therapy and 0.85 (0.81, 0.88) at the six-month time point.[288] The sensitivity and specificity of the shortened tool have been questioned.[205] The QuickDASH may underestimate symptoms and overestimate function when compared to the DASH.[285] QuickDASH Item 6: *recreational activities in which you take some impact* is the most difficult item for individuals undergoing total shoulder arthroplasty or rotator cuff repair, and Item 10: *tingling* is the easiest item.[224] The QuickDASH has been used in the HNC literature as a measure of shoulder function,[87] in a study looking at function of the upper extremity in the setting of ulnar forearm flap harvest.[289]

DASH & QuickDASH: Rasch Analysis

The DASH and the QuickDASH have been analyzed with Rasch methodology. While the majority of the analysis has occurred on the DASH, recent studies have focused on the QuickDASH.[10,283] Franchignoni and colleagues question the usefulness of the QuickDASH based on unresolved weaknesses found in their analysis of the DASH specific to misfit of two items, Item 26: *tingling* and Item 21: *sexual activity*, one of which, Item 21: *tingling*, is also used in the QuickDASH (Item 1).[10,14] Although both tools have demonstrated excellent psychometric properties under the CTT model across many populations, Rasch analysis suggests limitations in

construct validity related to step calibration and scale structure, scale hierarchy, item misfit, and violations of the assumptions of item interdependence and unidimensionality in patients with musculoskeletal complaints,[10,13-17] stroke,[18] and in patients with MS.[11] Three of these studies are specific to hand-related symptoms and will not be discussed further.[13,16,17] Preliminary Rasch analysis of the DASH and QuickDASH in 131 subjects undergoing neck dissection for treatment of HNC confirms weakness in the construct validity of the measures with similar limitations.[19]

The step calibration and scale structure for the DASH are described for a population of patients with MS,[11] general musculoskeletal complaints,[14] and patients with HNC.[19] The QuickDASH is also analyzed in the HNC population.[19] Disordered response option thresholds were found for 9-15 items on the DASH and 2-3 items on the QuickDASH, suggesting that the scoring system is not working as intended for the scales.[11,19,283] While specific information was not provided in one study consisting of subjects with upper extremity disorders, the authors report limitations for all items in response scale structure, specifically related to threshold separation or step calibration.[14] Each study reports underutilization of response options, with some of the studies attempting to collapse the response options for more accurate person discrimination.[10,13,14,19,283] The MS study found underutilization of the fourth response option ('severe difficulty'),[11] while the HNC study found underutilization of the fourth and fifth ('unable') response options.[19] The stroke study suggests the need to collapse categories but does not elaborate.[18]

The DASH does offer a good ability to measure test responders with MS across the ability continuum, however test items were redundant in the middle ability level and gaps were present at the lower ability levels.[11] Similar findings were identified in the HNC analysis, which demonstrated a good ability of both the DASH and QuickDASH to measure ability level at the

moderate to high disability levels and an inability to measure ability at the low end of the scale. The DASH was found to have nine areas of item redundancy, and two areas on the QuickDASH.[19] Ceiling effects are also reported for both scales.[19] In a population of 1030 individuals with shoulder pain, Rasch analysis of the QuickDASH shows good targeting of ability levels with some skew toward higher ability and gaps in the scale structure for the more able persons (easier items).[283]

Item misfit was found across patient populations. The greatest number of misfitting DASH items was found in the MS patient population, in which seven items demonstrated significant item misfit: Item 7: *heavy chores*, Item 11: *carry heavy objects*, Item 5: *open door*, Item 18: *recreational activities requiring force or impact*, Item 28: *stiffness*, Item 30: *feel less capable,* and Item 26: *tingling*.[11] In the HNC study, Item 26: *tingling*, Item 21: *sexual activities,* and Item 30: *feel less capable* misfit the model for the DASH.[19] Of these items, the only retained item in the QuickDASH, Item 10: *tingling*, consistently misfits the model.[10,19,283] Item 26: *tingling* and Item 21: *sexual activities* also misfit the model for the DASH in upper extremity musculoskeletal populations.[14] Lehman and colleagues note that Item 21: *sexual activities* and Item 20: *manage transportation* fail to load during exploratory factor analysis, and report item misfit for Item 29: *sleeping*, Item 26: *tingling* and Item 21: *sexual activities* in a similar population.[15] Dalton and colleagues found misfit with Item 21: *sexual activities* and Item 27: *weakness in arm, shoulder or hand* in a population of individuals with stroke, however eliminated the misfit issue with separate calibrations for purposed subscales of pain and impact scales.[18] QuickDASH Items 5: *use a knife to cut food* and 11: *sleeping* misfit the model in the large sample of individuals with shoulder pain. The authors suggest that when cutting food, the arm is usually held by the side and does not impact the shoulder. They question the sleeping item

as misfit because sleep impairment is a common complaint amongst individuals with shoulder impairment.[283]

The assumptions for IRT and Rasch analysis are unidimensionality and item interdependence. The unidimensionality of the DASH has been disputed. Factor analysis suggests multiple dimensions, however, analysis by the developing authors found sufficient evidence to support a single dimension and therefore inclusion of all 30 items within the composite score.[242] Lehman and colleagues performed exploratory and confirmatory factor analysis of the DASH and found three conceptually distinct groups– gross motor activities requiring whole-body movements, fine motor items, and symptom items. The authors were able to provide evidence to support the DASH as both a unidimensional tool and a multidimensional tool, however ultimately recommended utilizing the three-factor scale to assist with scale interpretation.[15] Dalton and colleagues also found three subscales: pain, impact, and function; however found that the pain and impact items can be combined appropriately, leaving a bi-dimensional scale.[18] The authors of the HNC study report a unidimensional scale with 2 distinct subdomains of symptoms and activities which require greater functional demand of the upper extremity.[19] On the other hand, other authors report evidence to support a multidimensional scale structure.[11,14] An analysis of the QuickDASH using parallel analysis and exploratory factor analysis supported a unidimensional scale with one misfitting item (Item 10: *tingling*) failing to load meaningfully.[10] However, Jerosch-Herold and colleagues did not find unidimensionality.[283] Rasch analysis also uncovers potential violation to the assumption of item interdependence. In the MS study, six item pairs were found to be highly correlated suggesting item dependency and ordering effect,[11] and 10 item pairs were found in the study assessing the DASH in general upper extremity diagnoses.[14] In the QuickDASH, two testlets have been recommended due to item

interdependence – 'household activities' (items 1, 2 and 3) and 'participation' (items 7 and 8). DIF is also reported based on age and sex.[283]

Because the DASH and QuickDASH are region-specific, rather than joint-specific, their specificity and responsiveness may be lower than other instruments that measure the shoulder joint only.[217,258] Medical management of HNC rarely affects function at the elbow, wrist and hand; therefore, the DASH and QuickDASH may not be specific enough to the limitations of activity and participation specific to this unique population. A potential strength of the measures is seen in recent research suggesting the DASH and QuickDASH may be appropriate for quantifying disability in patients with symptoms in the neck and upper extremity.[247,254,255] These studies may prove relevant to the management of HNC with neck dissection procedures and resulting upper extremity dysfunction.

Neck Dissection Impairment Index

The NDII, first published in 2002, is a 10-item PRO designed to quantify HRQOL secondary to shoulder impairment. Test items are scored on a 5-point Likert scale, with response options including 'not at all,' 'a little bit,' 'a moderate amount,' 'quite a bit,' and 'a lot.' A response of 'not at all' is given an item score of five points. The raw NDII score is scaled to a cumulative score of 100 using the equation: [(raw score – 10)/40] x 100. A higher score suggests higher QOL, and therefore less neck and shoulder impairment. Test item recall is four weeks.[4] Factors that contribute most to the total NDII test score include age, weight, type of neck dissection performed, and treatment with radiotherapy.[4] Techniques to manage missing data are not reported. The NDII has low responder burden, and takes only five minutes to complete.[3]

The measure was developed in a sample of 54 patients (64 neck dissections) - 32 of which were SAN-sparing MRND and 32 of which were SND procedures - who were an average

of 33.7 months (range, 11-120 months) from surgery. The sample is representative of the typical HNC patient, however patients who have had surgery for HNC within the past 11 months, have a known recurrent cancer at the time of the study, or an unrelated neck or shoulder condition were excluded. Most of the patients had previously undergone radiation therapy. Test items were derived from a review of the literature, patient interviews, expert opinion of otolaryngologists, physical therapists, and survey specialists, and pilot testing in a group of 25 patients undergoing neck dissection surgery. Item reduction eliminated five test items based on poor test-retest reliability, leaving 10 items addressing one domain related to physical abilities and activities.[4]

The population mean for all patients in the validation study is 67.8±17.4 (range, 7.5-100.0).[4] This is similar to the population mean found in Goldstein's study of 96 subjects aged 62.7 (range not provided) who underwent RND, SAN-sparing MRND, or SND greater than 11 months prior to enrollment who scored on average 74.3±25.79 (range, 2.5-100);[196] and significantly lower than the NDII score for subjects aged 57.28 (range, 46-70) years who are 18 months following bilateral neck dissection and total or partial laryngectomy or hemiglossectomy who scored a mean ± SD of 98.2 ± 1.98 (range, 95 – 100).[38] Wang and colleagues utilized the NDII to assess shoulder-related QOL in a cohort of individuals undergoing super-selective or SND for HNC after receiving chemoradiotherapy. In this population, subjects scored a mean 87.4 (SD, 12.3) prior to surgery, 80.2 (20.5) approximately 1.4 (range, 1-3.5) months after surgery, and 88.0 (13.8) 18 months (range, 10-37) months after surgery.[95] Gallagher and colleagues report a median (range) NDII score of 85 (30-100) following MRND (with dissection of level V) and 92 (10-100) following SND in a population of patients at least 12 month following neck dissection and completion of adjuvant treatments.[129]

Concurrent validity of the NDII was established with the SF-36 and criterion validity was supported through the establishment of convergent validity with the CS. The correlation between the NDII and the CS is 0.85[4] and 0.64.[129] The NDII demonstrates statistically significant correlation with the following SF-36 domains: physical functioning (0.50), role-physical functioning (0.60), role-emotional (0.59), social functioning (0.62), mental health (0.56), vitality (0.44), and general health perceptions (0.55). Bodily pain was not significantly correlated (0.32, $p = 0.005$).[4] Subsequent research has further established the convergent validity of the NDII through correlation with other shoulder-related measures, including the SDQ ($r = -0.77$),[290] shoulder subscale of the UW-QoL ($r = 0.75$),[290] and the DASH ($r = -0.86$).[196] Discriminative (clinical) validity of the NDII has been calculated using AUC of the ROC-curve with patient-reported need for physical therapy as the criterion. It was assumed that discriminative validity was confirmed if the questionnaire correctly classified patients with a self-reported need for shoulder therapy. At 1-3 months from surgery the AUC for the NDII was 0.85 (95% confidence interval, CI, 0.77-0.94). Six months later, the discriminative ability of the NDII is less, 0.74 (95% CI 0.58 – 0.90).[21] Rasch analysis of the NDII suggests limitations in the construct validity of the NDII secondary to disordered response option step categories, gaps in item difficulty, and item redundancy.[21] No ceiling effect, defined in the NDII as having the worst possible impairment of zero, has been reported. However, 5% of the sample scored the best possible score, a floor effect, of 100 at 1-3 months and 20% at 8 months following surgery.[21]

Reliability of the NDII has been established through test-retest reliability and internal consistency. Single-item ($r = 0.41$-1.00) and total score test-retest reliability ($r = 0.91$) and internal consistency ($\alpha = 0.95$) have been reported in the original validation study.[4] Stuiver and colleagues report similar reliability for a sample of patients 1-3 months from neck dissection

surgery with a test-retest reliability of $ICC_{(2,1)} = 0.93$ (95% CI, 0.87-0.96) and internal consistency of $\alpha = 0.94$.[21] Stuvier and colleagues suggest responsiveness to change based on a "visual assessment of change in mean scores" and "a strong association over time with a change of shoulder active ROM for abduction." Although not reported, the SEM and MDC can be calculated from the Stuvier paper (SEM = 5.56, $MDC_{95\%} = 4.67$).[21]

A limitation of the NDII is that it does not differentiate between neck and shoulder symptoms,[3,115] resulting in a risk of error in score interpretation. In addition, the sample utilized to establish reliability and validity only included those who had received a nerve-sparing procedure and therefore may not accurately reflect all individuals in the HNC population for whom the NDII will be provided.[3] The Stuiver study suggests limitations in the NDII's ability to accurately quantify subject ability level along the continuum of item difficulty levels, a limitation in the measure's construct validity.[21] The NDII has been used in HNC-related research,[4,21,38,71,115,128-130,195,196,239,290-292] however requires additional research to establish more robust psychometric properties including ICC for reliability, cross-cultural evaluation, responsiveness to change, normative values, and cut off scores to assist in interpretation of shoulder impairment in the HNC population.[3]

Shoulder Disability Questionnaire

The original version of the SDQ was developed in the United Kingdom (SDQ-UK) in 1994 as an "assessment of restriction in everyday activities resulting from shoulder symptoms."[293(p 525)] The 22-test items in the SDQ-UK were generated through the expert opinion of physical and occupational therapist, patient interviews, and item selection from a database of shoulder-related interview-based questions validated for use in the general population.[293] The SDQ-UK was rarely used,[205] and in 1998 was modified into the 16-item SDQ-Netherlands

(SDQ-NL).[294] The rationale for the modification and reduction of test items is not described, however. The SDQ-NL was developed in a cohort of patients with shoulder disorders in general practice with a minimum age of 18 years (mean 49.6, SD 14.4). Patients could not have neurological or vascular disorders, neoplasms, and referred pain from internal organs or systemic rheumatic conditions, fractures, or subluxations.[294] The 16 items in the SDQ-NL are right or left-shoulder specific and refer to the functional tasks performed in the preceding 24 hours.[294] Thirteen of the items relate to pain with activities and three items consider difficulty sleeping, the need to rub the shoulder, and irritability due to shoulder pain. The SDQ-NL was further modified in 2000. This version retained the same 16 items and scoring methodology and made only minor changes in wording based on expert opinion and patient interviews.[295]

When answering the questionnaire, the responder is required to answer 'yes,' 'no,' or 'not applicable' based upon if the situation has not occurred within the past 24 hours. 'Not applicable' test items, items that have not occurred, are excluded from scoring. To obtain a test score, the total number of 'yes' responses is divided by the number of applicable items, and then multiplied by 100 to provide a score on a scale of 0-100.[294] A score of zero is interpreted as no disability and a score of 100 is interpreted as maximum disability.[205] In the SDQ-NL, Items 7: *write or type*, 8: *hold steering wheel or bike handle bars*, 11: *open or close door* and 16: *irritability* are most often scored as 'not applicable,' meaning that the responder had not performed or experienced that activity within the past 24 hours.[295] The SDQ has minimal responder and tester burden, taking only three minutes to complete and one minute to score.[149,205] van der Heijden, however, reports a longer time (5-10 minutes) to complete the SDQ.[295] The SDQ is available in several languages, including English, Spanish, Korean and Turkish.[205,296-298]

Schmidt and colleagues suggest that the main limitation of the SDQ is the limited evidence to support the measure's reliability.[229] Reliability has been reported in populations of patients with shoulder pain, with internal consistency ranges from $\alpha = 0.76$ to 0.82,[297-299] and test-retest reliability has been reported as $r = 0.88$[297] and ICC (95% CI) of 0.79 (0.53-0.91).[298]

The validity of the SDQ has been questioned when compared with other shoulder-related outcome measures.[204] The SDQ has been reported as having a significant ceiling effect in subjects with soft tissue shoulder disorders, with an inability of the SDQ to distinguish between deteriorated and non-deteriorated subjects when test scores are relatively high at baseline.[204,295] In addition, low correlation coefficients comparing the SDQ to other instruments suggest limited convergent validity. Correlation coefficients have been reported as follows: VAS for function (0.58), VAS for pain (0.41), SPADI (0.33),[300] Korean SPADI (0.71-0.72) and Numeric Rating Scale (0.65-0.71),[298] and ROM (0.27-0.41).[205] In a population of patients presenting to primary and secondary care for shoulder pain, de Winter and colleagues propose content validity based on acceptable correlations with assessments of pain, ROM, strength, performance of ADLs, and disability quantified with a VAS. They also established content validity by comparing patient complaints with the items of the SDQ, and demonstrated the discriminative validity of the SDQ in the primary care setting, but not in the secondary care setting.[299] Of interest, in a comparison of the SDQ-NL, SDQ-UK, SPADI, and the Shoulder Rating Questionnaire in a population of patients presenting with shoulder pain in a primary care setting, subjects ranked the SDQ-NL as the most relevant of the four measures to their shoulder symptoms.[300]

Responsiveness to change for the SDQ has not been established. MDC, SEM, and MCID values are not reported,[3,205] however, ES and SRM for various orthopedic shoulder complaints are published in the literature.[205] In the primary care setting, an ES of 1.56 and a SRM of 0.95

are reported.[300] In a population of Korean patients with shoulder impairment an ES of 1.55 and SRM of 1.30 is reported.[298] Mean change scores for subjects who had improved are 29.2 points and 2.8 for those who remained stable.[300] A cut-off score of 18.75 balances sensitivity (74%) and specificity (66%).[301] The limited response options may limit the sensitivity of the measure in detecting change over time, or differences between groups, therefore limiting applicability in both clinical and research settings.[3]

The SDQ has been used in research related to SAN-injury[236] and HNC,[21,82,290,302,303] and relevant psychometric properties have been reported. In 2015, Cho and colleagues published a manuscript detailing the difference in trapezius muscle volume following neck dissection for HNC. The authors utilized a computed tomography (CT) scan and the original version of the SDQ, SDQ-UK, in their study. Because this review focuses on the revised SDQ-NL, additional details will not be provided.[304] Mean (SD) scores of the SDQ for neck dissection procedures have been reported as follows: posterolateral neck dissection (dissection of levels II-V) 48.6 points (35.1), MRND 22.2 (28.6), and SND 11.6 (26.1).[303] Normative values in another sample of patients with HNC reports a mean (SD) score of 33 (29) for those receiving neck dissection and 13 (27) for those who received non-surgical management via chemoradiation.[82] Subjects were most likely to endorse Item 10: *reach above shoulder level* (38%), Item 15: *rub the shoulder more than once per day* (34%), Item 4: *daily activities* (31%), Item 9: *lift or carry an object* (31%), Item 14: *reach the back of the neck* (29%), and Item 2: *lying on the affected shoulder* (29%).[303] VAS pain score, shoulder stiffness, ROM (abduction, flexion), receiving physical therapy, shoulder drooping, and surgical treatment of the neck were predictive of shoulder dysfunction (SDQ score > 0) as measured by the SDQ.[82]

Test-retest reliability is reported as $ICC_{(2,1)} = 0.84$ (95% CI, 0.74-0.90), and internal consistency as $\alpha = 0.91$.[21] Stuiver and colleagues found a significant floor effect in their sample of patients who had undergoing neck dissection procedure, with 18% scoring the best score (no limitation) at one to three months following surgery and 52% of the sample reporting no limitation six months later.[21]

Convergent validity of the SDQ has been demonstrated with correlation between the SDQ and the NDII (-0.77),[290] shoulder subscale of the UW-QoL (-0.75),[290] and the VAS for pain (0.631).[82] Discriminative validity of the SDQ has been supported using AUC of the ROC-curve with patient-reported need for physical therapy as the criterion. At 1-3 months from surgery the AUC for the SDQ was 0.85 (95% confidence interval, CI, 0.78-0.94) and 6 months later 0.77 (95% CI 0.63 – 0.91).[21] In other words, at 1-3 months from surgery the SDQ is able to accurately discriminate between those who believe they need physical therapy and those who do not 85% of the time. The discriminative validity decreases to 77% six months later.[178] There is evidence to support that the SDQ is unable to discriminate between neck dissection types, between subjects receiving adjuvant radiotherapy and those who are not, and those receiving primary radiation and those receiving chemoradiation.[82] Stuvier and colleagues suggest responsiveness to change based on a "visual assessment of change in mean scores" and "a strong association over time with a change of shoulder active ROM for abduction." Although not reported, the SEM and MDC for the SDQ can be calculated from the Stuvier paper (SEM = 11.36, $MDC_{95\%} = 31.49$).[21]

Shoulder Pain and Disability Index

The SPADI was first published in 1991 by Roach and colleagues as a patient-reported measure of shoulder pain and disability. The SPADI was developed through the generation of 20 items based on the expertise of rheumatologists and physical therapists. Seven test items were

then eliminated based on low test-retest reliability and correlation with active shoulder ROM measurements in a sample of 23 male patients with shoulder pain of musculoskeletal, neurogenic or undetermined origin that were on average 58 years old (range 23-76).[305] The measure therefore consists of 13 items that assess the two dimensions (subscales) of pain (five items) and functional activities that require the use of the upper extremity (eight items).[305]

There are two versions of the SPADI, which can be used interchangeably.[306] The original version utilizes a VAS with anchors of 'no pain' or 'no difficulty,' and 'worst pain imaginable' or 'so difficult required help.' Test items are scored by "arbitrarily dividing the horizontal line into 12 segments of equal length," each representing a score of 0-11.[305] On the second version the responder answers using an 11-point Likert scale in which a score of zero indicates 'no pain' or 'no difficulty' and a score of 10 indicates 'the worst pain imaginable' or 'so difficult it requires help.'[307] The recall period is one week. Scoring provides a total pain score, a total disability score, and a total SPADI score, which are each expressed as a percentage of 0-100%. The total SPADI score is the mean of the pain and disability scores. Responders can mark one item in each subscale as not applicable and that item is subsequently removed from scoring. No score can be calculated if more than two items are missing or marked as not applicable. A score of 100% indicates the highest level of impairment or disability.[305] Cut off points to reflect severity of disability are not reported.[205] The SPADI takes 5-10 minutes to complete, and is scored in 1-8 minutes (based on which version of the scale is used).[149,305,306,308] The SPADI is widely used across multiple patient populations, including general upper extremity diagnoses, general shoulder diagnoses, adhesive capsulitis, rotator cuff disorder, and shoulder surgery.[205] It has also been utilized in a cross-cultural validation study using a population of women within six

months of breast cancer treatment.[309] In addition, it has been translated and validated in several languages.[194,202,205,216,234,268,298,309-313]

The SPADI has been shown to have good reproducibility. Test-retest reliability (ICC) ranges from 0.66 to 0.95[3,149,234,256,265,305,306] and internal consistency (α) ranges between 0.86 and 0.98.[3,149,265,305,306,309,314] Internal consistency (α) for the pain subscale is reported as 0.85 and the disability pain scale is 0.90.[315] On the Chinese version, the ICC is 0.836 for both scales.[234] The original study reported a test-retest reliability of ICC = 0.66. In this study, test-retest reliability was assessed based upon responses obtained within a 48-hour period, during which 91% of the subjects began interventions such as corticosteroid injections and pain medications. The accuracy of the test-retest reliability statistic should be questioned because the researchers did not provide evidence to suggest elimination of subjects based on reported change in symptoms during that timeframe.[305] Membrilla-Mesa and colleagues report a test-retest reliability statistic of the Spanish version of the SPADI as r = 0.89-0.93,[268] and the Chinese version is reported as ICC 0.85-0.90.[234] SEM (95% CI) for the SPADI pain subscale is reported as 7.82 ± 15.3, disability subscale 5.78 ±11.3 and for the total score 23.8 ± 9.3. The high SEM and confidence intervals suggest that the SPADI, while appropriate for group level research, may not be precise enough on an individual level.[314] Roy and colleagues performed a review of shoulder-related questionnaires, and report a SEM of 6.2 to 7.8 (mean 6.8 points).[256,265] In a population of women undergoing breast cancer treatment, an ES of 0.59 and SRM of 0.78 on the disability scale and an ES of 0.82 and SRM of 1.13 on the pain subscale.[309]

The convergent validity of the SPADI has been well established through assessment of the scale's correlation with other measures of shoulder function, with correlation coefficients as follows: pASES (0.81-0.92)[3,202,205,216]; CS (0.69 and 0.82)[216,234]; DASH (0.55-0.93)[90,205,216,268];

SST (-0.71-0.80)[205,268]; SDQ-K (Korean version, 0.71-0.72)[298]; Oxford Shoulder Scale (0.674)[309]; and VAS (0.40 – 0.67).[234,258,268] The SDQ-NL has a fair correlation with the SPADI, whereas the SDQ-UK is moderately correlated (0.573).[300] Correlation with active shoulder ROM measurements for the SPADI have been reported in two studies, with correlations for various shoulder movements ranging from -0.55 to -0.80,[305] and -0.090 to -0.251.[300] The tendency toward low correlations with active ROM has led to some discussion regarding the development and item reduction of the SPADI based on low correlations with active ROM measurements.[204] The SPADI has also demonstrated discriminant validity.[3] Williams and colleagues report a AUC of 0.91, where a AUC of 1.0 indicates 100% accuracy in ability to discriminate between health states.[307] Thoomes-de Graaf report and AUC of 0.81 (95% CI, 0.75-0.87).[316] Similar findings are reported in a sample of individuals undergoing physical therapy for shoulder pain. At the 6-week follow-up the AUC is reported as 0.81 (0.78, 0.84) and at the 6-month follow-up 0.85 (0.82, 0.88).[288] The SPADI has been found to have no to low floor or ceiling effects for the total score,[204,205,234,265,306,309,312,313] however, the pain subscale did demonstrate ceiling effect in patients receiving total shoulder arthroplasty,[216] and several items on the pain and disability subscales were reported to have relatively high ceiling effects in a Chinese population of patients with chronic shoulder pain.[234]

Scale developers intended to develop a 2-dimensional scale, which was subsequently supported in other studies[234,268]; however factor analysis and other psychometric studies have provided evidence to suggest that the pain and function subscales of the SPADI represent only one dimension.[305,314] It can be suggested that the construct of pain is impacting how responders are answering these function-based test items. Roddey and colleagues postulate that responders do not differentiate between pain and disability when responding to test items, citing the

following findings: (1) a high internal consistency ($\alpha = 0.96$); (2) principal-component factor analysis resulted in a one-factor solution; and (3) the subscales of the SPADI (pain and function) are highly correlated ($r = 0.77$).[314] The necessity of having two subscales within the SPADI has therefore been questioned.[314]

Rasch analysis of the SPADI has uncovered unequal interval measures, and poor precision measuring shoulder dysfunction at the low and high ends of the scale. In fact, the SPADI performed worse than the ASES and the Penn Shoulder Scale (function subscale) for capturing subjects with low and high levels of disability. The SPADI did however perform better than the ASES, the Penn Shoulder Scale, and the SST, in measuring middle range scores. This imprecision at the high and low ends of the scale negatively impacts the ability of the scale to measure the construct of shoulder dysfunction at low and high ability levels, therefore decreasing the construct validity of the measure.[20] Thoomes-de Graaf supports this finding of poor precision at the high and low ends of ability, where only 2.2% of the sample scored in the upper range (85-100%) and 8.1% scored in the low range 0-15%).[312] Rasch analysis has also uncovered some concerns with item misfit for Item 13: *removing something from back pocket*, Item 12: *carrying a heavy object of 10 pounds or more*, and Item 9: *putting on a shirt that buttons down the front* in a population of individuals with shoulder impairment.[20]

The SPADI's responsiveness to change in the general musculoskeletal populations has also been established. $MDC_{90\%}$ is reported as 13-18 points,[256,305] and an $MDC_{95\%}$ ranges from 13.2 to 21.5 points.[205] MCID scores range from 8 to 23.1 points based on the sample.[205,300] Thoomes-de Graaf and colleagues report a limitation in interpretability of the SPADI in a population of patients with should complaints presenting to a primary care physical therapy setting. Although a minimal important change, the smallest change that patients perceive as

important, is reported as 16 points, the authors suggest that a change score of less than 20 points could be due to measurement error. According to the authors, a change score of 43% in individuals is clinically relevant.[312] Whereas, other studies indicate an improvement of 10 points on the SPADI is indicative of a 12% improvement in shoulder function (likelihood ratio = 34), and a worsening of 10 points on the SPADI is indicative of a 31% decline in function (likelihood ratio = 12.9).[307] ES and SRM for various orthopedic patient populations have been reported.[205,258,265] Of these, the most relevant to the HNC population include: patients referred to outpatient physical therapy clinics for a wide variety of shoulder impairments (ES 1.26, SRM 1.38)[308]; patients attending a walk-in clinic at two large medical centers with complaints of shoulder discomfort (ES 0.34)[307]; and a population of patients presenting to a primary care setting with shoulder complaints (ES of 1.52, SRM of 1.17).[300]

The SPADI has been used in HNC-related research[21,79,109,114,115,128,194,195,199] and has been translated into an Italian version and specifically validated in a population of patients with HNC.[194] Swisher utilized the SPADI to study QOL and shoulder impairment in a sample of 37 subjects [84% male, 63.7 (SD, 11.4; range, 30-82) years old] following surgery, radiation, and/or chemotherapy for HNC. Normative values were established with a mean (SD) total, pain, and disability scores for the SPADI reported as 19.71% (23.70%), 23.89% (27.72%), and 18.59% (23.10%), respectively. According to the authors, the subjects had the greatest difficulty with the following items: Item 7: *washing the back*, Item 11: *placing an object on a high shelf*, and Item 12: *carrying a heavy object*.[109] A significant floor effect has been found in a population of patients with HNC who are 1-3 months following surgery with 17% of subjects reporting no disability. A larger floor effect of 43% was found 6-8 months following surgery.[302] Ghiam and

colleagues utilized a cut off score of greater than 30 on the SPADI subscales to define high levels of shoulder pain and disability.[199]

The SPADI has been found to be reliable and valid in the HNC population. Test-retest reliability is reported as $ICC_{(2,1)} = 0.91$ (95% CI, 0.85-0.95), and internal consistency of $\alpha = 0.96$.[21] Convergent validity was established with the SPADI and the Neck Disability Index ($r = 0.86$) and the composite score of the UW-QoL ($r = -0.73$).[109] The SPADI also has good convergent validity with other measures of shoulder function in the HNC population, including the NDII ($r = -0.75$) and the SDQ ($r= 0.78$), and the Rand-36 (SF-36) in which correlations ranged from -0.19 to -0.55.[21] Discriminative validity of the SPADI has been supported using AUC of the ROC-curve with patient-reported need for physical therapy as the criterion. At 1-3 months from surgery the AUC for the SPADI was 0.85 (95% confidence interval, CI, 0.77-0.94) and six months later 0.74 (95% CI 0.58 – 0.90).[21] Rasch analysis of the SPADI in a population of patients with HNC, however, suggests limitations in the construct validity of the scale secondary to disordered response option step categories, gaps in item difficulty, and item redundancy.[21] Stuvier and colleagues suggest responsiveness to change based on a "visual assessment of change in mean scores" and "a strong association over time with a change of shoulder active ROM for abduction." Although not reported, the SEM and MDC for the SPADI can be calculated from the Stuvier paper ($SEM = 6.54$, $MDC_{95\%} = 18.13$).[21]

Simple Shoulder Test

The SST was developed by the Shoulder Service at the University of Washington with questions obtained from Neer's evaluation, the ASES evaluation, and observations of patient complaints by the test developers.[200,205,317] It includes 12 dichotomous 'yes' or 'no' items which address the domain of function.[149] Two items are related to pain, seven are related to function

and strength, and three items are related to ROM.[318] The recall period is at the moment of assessment. The test items are scored one point for an answer of 'yes' and zero points for an answer of 'no.' The test score is based on a 0-12 point range, which is then converted to a percentage. A 100% indicates no shoulder limitation. The SST takes three minutes to complete and one minute to score.[149] The SST has been translated into an Italian version and specifically validated in a population of patients with HNC.[194] It is also available in Dutch,[319] Brazilian Portuguese,[320] Persian,[321,322] and Spanish.[323] A study of young, active adults, mean age 28.8 (range 17-50 years), scored an average score of 11.79 points (SD, 0.60) on the SST.[220] Pre-operatively for reverse total shoulder arthroplasty or rotator cuff repair, individuals aged 25-89 years scored 2.30 (SD, 2.31; range, 0-11); and post-operatively, individuals aged 37-88 scored 6.10 (SD, 3.15; range, 0-12).[324]

The SST has been utilized in populations with various shoulder conditions including impingement,[318] OA,[318,320] and instability,[318,325] rotator cuff injuries,[225,320,325] rotator cuff surgery, and total shoulder arthroplasty.[221,324,326] It has also been utilized in general orthopedic practice.[314] The SST has not been used in interventional studies related to HNC with the exception of the cross-cultural adaptation and validation of the Italian version of the scale.[194] Test-retest reliability has been reported as ICC = 0.61-0.99[149,221,320-323,325] and internal consistency has been reported as (α) 0.73 - 0.85[314,320,322] Person reliability, consistent with Cronbach's α, is reported in a study that used partial credit Rasch analysis as 0.66[20] and 0.71.[225] An SEM of 11.65 and an $MDC_{95\%}$ of 32.3 points have been reported.[314] A systematic review of PROs for individuals presenting with rotator cuff disorder reports an MDC of 3.27 and a minimal important difference (MID) of 2.05.[315]

Convergent validity of the SST has been supported through established relationships with other shoulder-related PROs as follows: DASH (- 0.596 to -0.73),[322,323,327] SPADI (- 0.71 to - 0.80),[268,314,318] ASES (0.54-0.81),[209,225,318,325] CS (0.49 – 0.70),[226,327] Patient Reported Outcomes Measurement Information System (PROMIS) Physical Function Computerized Adaptive Test (0.64),[225] and the SF-36 physical functioning scale (r = 0.47 - 0.58), physical component score (r = 0.54) and bodily pain scale (r = 0.57 - 0.62).[318,326] ES and SRM for several orthopedic populations are also reported in the literature.[205,326] Low to no floor and ceiling effects have been reported,[205,321] however Hsu and colleagues report a floor effect in 9% of their sample (individuals receiving total shoulder arthroplasty) and a ceiling effect for 15.3%.[326] A floor effect of 21% and ceiling effect of 6.1% is reported in a sample of 187 individuals with rotator cuff pathology.[225]

There is discrepancy in the literature related to dimensionality of the SST. Beckmann reports that the SST is "largely unidimensional," where 8.4% unexplained variance remains after the first dimension is accounted for.[225] Roddey and colleagues found a 2-factor solution when performing principal-components factor analysis on the SST, suggesting domains related to 'what a person can *do* with his or her shoulder' and 'a person's *comfort* with the shoulder at rest.'[314(p 766)] Neto and colleagues found three well-defined factors through factor analysis: 'arm elevation,' 'shoulder movement,' and 'comfort with the shoulder in rest position.'[320] Several other studies have confirmed a three-factor solution for the SST.[321,323]

The SST has been tested using Rasch analysis.[20,225,324] In a sample of 187 individuals with rotator cuff pathology, item reliability is reported as 0.97 and person reliability as 0.71.[225] Rasch analysis has suggested problems in item misfit. In 2017, Raman and colleagues published a study of a population of individuals prior to reverse total shoulder arthroplasty or rotator cuff

repair and another group of individuals 6-12 months following surgery using Rasch analysis. Analysis showed misfit of three items: Item 4: *Can you place your hand behind your head with the elbow straight out to the side?*, Item 5: *Can you place a coin on a shelf at the level of your shoulder without bending your elbow?,* and Item 8: *Can you carry 20 pounds at your side with the affected extremity?*. Item 8 was found to have DIF based upon gender. In addition, local dependency was found for Items 4 and 5 and 5 and 6: *Can you lift on pound (full pint container) to the level of your shoulder without bending your elbow?*. The authors suggest combining items 5 and 6 to create a super item that addresses lifting an unspecified weight to shoulder level, and splitting Item 8 to a female and male item to address these limitations. With these modifications, the authors claim that misfit is negated.[324] Another study using Rasch analysis found the SST to have two different items that misfit the model with high infit statistics - Item 2: *Does your shoulder allow you to sleep comfortably?* and Item 1: *Is your shoulder comfortable with your arm at rest by your side?*. Rasch analysis found a failure of the SST to adequately separate subjects at the high and low end of the scale. The analysis suggests that the SST does not have equal interval measures, suggesting a limitation in responsiveness to change over time.[20]

Limitations of the SST include limited responsiveness to change and discriminative validity due to the dichotomous response options, lack of normative data, large floor effect, and moderate person reliability. Another limitation is the inclusion of test items to require the responder to speculate on their ability to perform a task.[3,205,225] For example, item 9 asks "Do you think you can toss a softball under-hand twenty yards with the affected extremity?"

University of Washington Quality of Life Questionnaire

The UW-QoL was first published in 1993 as a QOL instrument applicable to "the broad group of head and neck cancer patients encountered at any institution or in multi-institutional

trials."[328(p 487)] The PRO has since become one of the most commonly used QOL scales in HNC research and is often selected to establish convergent validity with other measures of QOL.[5,301,329,330] The original version consists of nine categories (pain, disfigurement, activity, recreation/entertainment, employment, eating-chewing, eating-swallowing, speech, and shoulder disability), each of which has 4-5 response options. Each category has a total possible point value of 100, and therefore a total summary score of 0-900 points. A higher score indicates "normal function" and higher QOL.[328] Information regarding test development has not been published.[5,331] According to Pusic and colleagues, personal communication with the developers of the UW-QoL reveals that scale development was based upon expert opinion and did not include patient interviews.[5]

The UW-QoL has subsequently undergone several major revisions based on shortcomings observed in the scale. Published in 1997, Version 2 added an importance rating scale, which requires the responder to rank the importance of each domain ('not important', 'a little bit important', 'somewhat important', 'quite important', or 'extremely important'), three single-item QOL scales, and a free-text opportunity in which the responder can elaborate on any medical or non-medical condition impacting perceived QOL.[332] Weymuller and colleagues describe a limitation in the concurrent validity of the UW-QoL composite score – the sum of each of the domains weighted for level of importance placed by the responder - with the addition of the importance rating scale in 1995. The authors also suggest a limitation in the sensitivity of the composite score based on the tendency of specific domains within the composite score to improve while others worsen during treatment and as a result of variations in treatment. The summary score may therefore not provide an accurate interpretation of QOL and responsiveness to change.[333] Version 3, also known as the UW-QOL-R, was published in 2001. Upon

consideration of internal consistency across domains, the authors of the scale removed the employment domain, and added the taste and saliva domains. Internal consistency for Version 3 ranges from $\alpha = 0.74\text{-}0.84$. Version 3 also includes a modification of the importance rating from ranking each of the domains to selecting the three most important domains in the past seven days.[334] The most current version of the UW-QoL, version 4 (UW-QoLv4), was published in 2002 with the addition of domains for mood and anxiety.[335] A 2012 paper published by Ghazali, Lowe, and Rogers suggests using an additional item to specify whether each domain had worsened, stayed the same, or improved over the last month. The authors suggest that his modification to the UW-QoLv4 may be of benefit in directing need for intervention.[336]

The UW-QoLv4 contains 12 domains (pain, appearance, activity, recreation, swallowing, chewing, speech, shoulder, taste, saliva, mood, and anxiety), each with 3-6 response options. Scoring is achieved in the same way as the original version, using a scale of 0-100, with a composite score of 0-1200. UW-QoLv4 retains the three items that address HRQOL and overall QOL, each scored on a scale of 0-100, the importance rating scale, and the free-text item .[335] Based upon limitations described by Weymuller and colleagues in 2000, calculation of an overall summary score for the UW-QoL is not recommended.[333,337] In 2010, Rogers and colleagues published evidence to support the presence of two UW-QoLv4 subscales based on the results of factor analysis. The physical function subscale includes the chewing, swallowing, speech, taste, saliva, and appearance domains, and the social-emotional subscale includes the anxiety, mood, pain, activity, recreation, and shoulder function domains. Rogers and colleagues suggest that the wording of the shoulder function scale reflects more on the domain of work and hobbies, causing it to load on the social-emotional domain rather than the physical function domain. Based on the finding that specific domains load to each subscale, a summary score (a simple average of each

item score, scale of 0-100) can be calculated for each subscale. Scoring requires that at least four items are answered in the subscale.[337]

Normative values for 372 subjects without HNC are 95 ± 10 (mean ± SD) for physical function and 83 ± 19 for social-emotional function. In a sample of 517 patients 1-2 years after surgery for HNC, the physical function score is 71 ± 21, and the social-emotional function score is 74 ± 20. Normative values based on cancer staging, presence of surgical flap, and radiotherapy are also available. Older patients exhibit a weak tendency to report better scores for both subscales, with no significant difference between genders. A mean change score of four units in a subscale score is interpreted as a small change, 10 units for a moderate change, and 16 units for a large change. For patients with baseline data prior to surgery, a mean change score of three units would indicate a small change, and 7.5 units and 12 units a moderate and large change, respectively.[337]

The original UW-QoL's psychometric properties were established in a population of 75 patients with HNC (69% male, mean age 55 years, range 23-83) through a comparison of two other QOL questionnaires, the Sickness Impact Profile (SIP) and the Karnofsky Performance scale, and are reported as 0.82-0.96 and 0.79-0.85 respectively.[328] Convergent validity has been established with other measures of QOL for patients with HNC, including the European Organisation for Research and Treatment of Cancer (EORTC) coreQOQ-C30,[335,338] the EORTC Head and Neck (H&N35),[338] and the SF-36.[338] Discriminant validity is supported based on the findings that individuals without HNC score significant higher on the UW-QoL than those with HNC.[339] No notable floor or ceiling effects have been observed.[340] The UW-QoL has not yet undergone psychometric testing using IRT or Rasch analysis, which would further strengthen the construct validity of the measure.

Internal consistency has been reported as $\alpha = 0.86$ for the 12-domain composite score, with the shoulder domain demonstrating the least correlation with the other 11 domains.[335] Test-retest reliability for the original version exceeded 0.94,[328] and for Version 4 the test-retest reliability for the physical function subscale ICC is 0.86, and 0.81 for the social-emotional function subscale.[337] Test item recall is seven days.[335] The UW-QoLv4 has been translated into 22 additional languages.[341] The Chinese version is a 13-item measure due to the addition of the previously utilized employment test item.[342] The tool is quick to complete (less than seven minutes) and easy to administer.[5,343] Scoring instructions are available, however, scoring is not intuitive.[328,340]

Because the UW-QoL is a multidimensional tool, the summary score may not be appropriately utilized to interpret impairment related to specific domains, such as shoulder dysfunction. As a result, some researchers have considered the utility of the shoulder question independent of the remainder of the UW-QoL in objectively quantifying shoulder impairment.[73,83,290] The shoulder subscale of the UW-QoL is scored on a scale of 0-100 points, and has four response options: 'I have no problems with my shoulder' (100 points), 'my shoulder is stiff but it has not affected my activity or strength' (70 points), 'pain or weakness in my shoulder has caused me to change my work/hobbies" (30 points), and 'I cannot work or do my hobbies due to problems with my shoulder' (0 points).[328] The subscale has remained unchanged across each of the 4 versions of the UW-QoL.

Rogers and colleagues compared the relationship between the shoulder subscale of the UW-QoL, SDQ, and NDII in a sample of 100 patients (54% male, median age 61 years, range 54-68) who were an average of 12 (range 3-38) months following surgery. The authors found that the subscale was strongly correlated with both the SDQ and the NDII (-0.75 for both

scales).[290] It should be pointed out that the scoring and interpretation of the NDII in the Rogers article varies from what was originally reported by Taylor and colleagues in the development and validation of the NDII. Rogers defines a higher score as greater impairment, whereas Taylor uses the opposite interpretation. As a result, the negative correlation of r = -0.75 for the NDII and the UW-QoL should actually be r = 0.75 because a high score in both measures indicates lower impairment of QOL.[4,290] Parikh and colleagues utilize the same interpretation as Rogers.[290,292] Nonetheless, results of the study demonstrate that the shoulder domain of the UW-QoL has positive predictive value for shoulder impairment, as determined by the NDII and the SDQ, when a score of lower than 100 points is reported, suggesting that the shoulder domain is an appropriate tool to screen for shoulder impairment in the HNC population.[290] A lower score on the shoulder domain also has positive predictive value for the occurrence of myofascial pain syndrome.[344] The shoulder domain also demonstrates strong correlations (r) with the following measures of QOL: SF-36 domains of role limitation physical (0.80), physical functioning (0.60), and pain (0.59); the UW-QoLv1 domains of activity (0.78), recreation (0.77), chewing (0.62), and disfigurement (0.62); EORTC C30 (+3) domain of role functioning new (0.59) and pain (0.58); and EORTC H&N35 domains of social contact with friends (0.62), pain in mouth (0.62), problems with teeth (0.57), and sticky saliva (0.56).[338]

In 2009, Rogers and colleagues published single-domain cut-off scores to trigger the need for intervention. The cut-off scores were based on the desire to limit the trigger for intervention to 1 in 5 patients. For the shoulder domain, the cut-off is 30 points and/or selecting the shoulder as an important domain in QOL. The authors postulate that the suggested cut-off score can serve as an effective screen for patients with shoulder impairment.[345]

Kuntz and Weymuller supported the use of the shoulder domain in evaluating shoulder function after neck dissection procedures in 1999. The authors performed a descriptive study of QOL, as measured by the UW-QoL, in 149 consecutive patients receiving RND, MRND, and SND procedures at baseline, six, and 12 months following surgery. The authors found that the shoulder domain adequately described the patient population and was able to differentiate shoulder dysfunction between the three neck dissection procedures.[83] Parikh and colleagues also consider the shoulder function domain independent of the physical function and social-emotional subscales in a surgery-based RCT.[292] In 2015, Garzaro and colleagues utilized the UW-QoL to compare shoulder function in individuals six months after neck dissection with sacrifice or sparing of the cervical plexus. The authors were able to discriminate between the two groups using the shoulder scale alone, finding a statistically significant worsening of function in individuals in which the plexus was sacrificed.[53]

Despite its increasing support in HNC literature, some authors suggest that the single item scale lacks "detail" and therefore has limited "usefulness in evaluating changes in shoulder complaints over time in the context of prevention and treatment trials."[21] Laverick and colleagues utilized the UW-QoL in a study to address QOL issues in patients receiving neck dissection surgery. As part of their analysis they individually considered the shoulder domain. The shoulder domain responded similar to the other domains of the UW-QoL scale across time points, however, the authors suggest that the 4-item scale offers a "crude assessment of shoulder function," and suggest that other shoulder-specific questionnaires may be more responsive, especially in the case of bilateral neck dissection surgeries.[73]

The UW-QoLv4 does have some limitations. The original scale was developed based on expert opinion with no published evidence to support item generation and reduction using sound

methodologies.[331] However, the extensive and robust psychometric studies that have been performed across the four revisions lend excellent evidence to support its validity and utility for assessing QOL in the HNC population. The appropriateness of the single domain shoulder function however remains to be determined. Another limitation of the UW-QoL reported in the literature is the scales applicability to individuals who have undergone surgery, rather than those receiving radiotherapy.[343] Rampling suggests added domains of voice and tiredness to better target individuals receiving radiotherapy.[343] Further research is needed to strengthen the construct validity of the UW-QoL, and to determine if the shoulder function domain should be used independently; however given the single-item nature of the UW-QoL (shoulder subscale) it cannot be analyzed using the partial credit Rasch methodologies employed in this study.

Rationale for PRO Inclusion in the Research Study

Of the shoulder-related PROs reviewed by Goldstein,[3] and recommended by the PULA Task Force[190] and the APTA EDGE Task Force,[2] the NDII and the UW-QoL (shoulder subscale) are the only measures which were specifically developed for the HNC population. The SPADI, SDQ, CS, ASES, SST, and DASH have been used in the HNC literature, however were originally developed for orthopedic and rheumatologic conditions. The QuickDASH has also been recommended despite its lack of use in the HNC population at this time.[2] Each of these measures has psychometric properties derived through CTT methodology, which are well documented in the literature through literature and systematic reviews.[149,200,204,205,228,229,235,265,306,315,346]

The lack of extensive research in this patient population related to shoulder-related PROs makes the development and comparison to other measures difficult and confusing. Recommendations are often clouded by personal researcher bias and incomplete information

regarding a PRO's psychometric properties. The DASH is frequently recommended over other shoulder-related outcome measures for use in multiple patient populations, including those with oncology diagnoses.[2,3,190,204,347] In a systematic review, Roy and colleagues recommend the use of the DASH and the ASES over the SPADI and the SST in the clinic and research settings when various shoulder disorders are evaluated. The ASES and SPADI are recommended if shoulder pain and physical function are being assessed. The ASES, however, cannot be used to provide a health utility score,[348] therefore the DASH is recommended if emotional and social function are of interest.[265] The DASH has been found to correlate highly with generic measures such as the SF-36, and may be able to decrease responder burden when a generic health utility score is needed.[349] Angst and colleagues recommend that the QuickDASH, along with the SPADI, pASES, and CS, be used for clinical use and the DASH, along with the SPADI, cASES, and CS, be used for research.[205] The NDII and SPADI demonstrate comparable reliability, while the SPADI provides more detail related to pain, the NDII provides more detail on more difficult items required for activity and social participation. The NDII has demonstrated superior sensitivity to type of neck dissection surgery received, and demonstrates the lowest floor effects when compared to the SPADI and the SDQ.[21] Thoomes-Graaf completed a systematic review of the literature with the intent to critically appraise and compare the measurement properties of the original and translated versions of PROs, which address shoulder-related activity limitations in individuals with non-specific shoulder pain. For the English language, the authors recommend the use of the SPADI due to its superior psychometric properties. The authors however suggest that their recommendations could be biased based upon the strict exclusion criteria. For example, many studies of the DASH or QuickDASH were excluded from their review because they reported results for individuals with elbow and/or hand impairment.[346]

In 2001 the ICF switched the focus from the medical disablement model to a model that classifies function and health through impairments in body function and structure, activity limitations and participation restrictions, all in the setting of environmental and personal factors.[89] In 2013, Roe and colleagues performed a systematic review of 40 shoulder-related outcome measures to identify which aspects of functioning were most frequently addressed. The authors found the concepts of pain, mobility related body functions and structures, sleep, hand and arm use, employment, recreation and leisure, and self-care to be the most frequently addressed. The authors found that the DASH and the ASES addressed the greatest number of ICF categories, nearly twice as many as the CS, SST and SPADI.[350]

The intent of this research study is to assess the appropriateness of the recommendations provided by the APTA EDGE Task Force for HNC.[2] It is recognized that this will be an incomplete analysis of the shoulder-related measures available to medical providers, but it will offer a framework from which to continue future research. For completeness, all shoulder-related PROs utilized in the HNC have been reviewed here.[3] Rationale as to why additional measures are not included in this study were considered. The CS includes a clinician-rated component that is included in the test score, and therefore does not constitute a true PRO. The SST and the SDQ both include dichotomous response options and therefore represent different scale structures than the measures recommended by the Task Force. The ASES is the most similar to the Task Force's recommended measures, however requires the responder to answer the questions separately for the right and left upper extremity. This methodology varies from the other recommended outcome measures and was therefore not included in the study.

Summary of What is Known and Unknown About PROs in the HNC Patient Population

With the greater than 50 shoulder-related outcome measures found in the literature,[2] the need for additional shoulder-related measures has been refuted based on the understanding that there will never be a "perfect" measure, and that too many measures will only confuse and hinder comparison across interventions or in research.[149,329,331] In 2007, Rogers and colleagues published a thematic review of HNC-related literature published on QOL and called for "ongoing validation to improve our understanding of existing questionnaires and also to develop subsite or function specific measures."[329(p 861)] Subsequently, Rogers published a similar paper in 2016 for literature published between 2006 and 2013 and called for "agreement about a specific group of functional outcomes" to "enable data to be pooled and compared across units" and to "improve the quality of data used to inform patients and the multiprofessional team about likely outcomes."[351(p e47)] When considering the unique presentation of patients presenting with SAN palsy and trapezius muscle atrophy in the setting of HNC, scales not originally intended for the HNC population, such as the DASH, QuickDASH and SPADI, may prove inappropriate.

Despite strong reported psychometric properties for the DASH, QuickDASH, and the SPADI across various patient populations, appropriateness of their use in the HNC population has not been fully supported. Rasch analysis of the DASH in various musculoskeletal and MS patient populations has uncovered limitations in construct validity related to item fit, dimensionality, item response option thresholds, response scale structure, and residual correlations.[10,11,13,14] The usefulness of the QuickDASH has also been questioned based on unresolved weaknesses found in the analysis of the DASH specific to misfit of two items, Item 26: *tingling* and Item 21: *sexual activity*, one of which, tingling (Item 10), was retained in the QuickDASH.[10,14] The SPADI, when tested with Rasch analysis in a population of patients with

general orthopedic complaints, was also found to have item misfit and poor precision at the low and high ends of the scale.[20]

Preliminary studies have utilized Rasch analysis to study PRO functionality in the HNC patient population. One study analyzed the DASH and QuickDASH, and found minor item misfit and ceiling effects, in addition to a large weakness in scale precision at the mild to moderate disability levels, suggesting a limitation in the scales' ability to discriminate between ability levels and demonstrate change.[19] Another study utilized Rasch analysis to analyze the psychometric properties of a measure that combines the SPADI and the NDII in a population of subjects following neck dissection surgery.[21] The analysis supported the unidimensionality of the combined scales, but showed disordered response scale structure, gaps in item difficulty coverage, and redundancies. The authors were able to resolve the disordered response options with dichotomization of response options on both scales, and were able decrease the gaps in item difficulty by combining the two scales.[21] The psychometric properties of the NDII has not been assessed using Rasch analysis at this time. Utilization of the single-item UW-QoL (shoulder subscale) does not allow for analysis using Rasch methodologies, however data will be included to allow for comparison of its summary score to the other PROs selected. This will allow for further consideration of the item's usability as a screening tool.

Summary of the Chapter

Survivors of HNC are often left with disfiguring impairments in body function and structure, resulting in activity limitations, participation restrictions, and overall disability. Shoulder dysfunction in the setting of SAN palsy and trapezius muscle atrophy is a common complaint among individuals who undergo a surgical neck dissection procedure. Although physical activity and exercise are recommended for all cancer survivors, HNC survivors tend to

be sedentary before, during and after cancer treatments. HNC survivors are infrequently referred to physical therapy to address functional deficits related to pain, decreased ROM and strength.

Progression toward reimbursement to pay for performance models requires physical therapists to demonstrate and document value for the care that is provided. Value is determined by the cost of the intervention divided by the outcome. Outcome measures, that are reliable and that have demonstrated validity in the population of interest, are necessary to show value. The inability of physical therapists to show the value of the interventions provided may have serious financial consequences related to the ability of HNC survivors to receive skilled physical therapy services, and for clinicians to be reimbursed for the care provided.

Although there are many PROs available to the physical therapist to quantify shoulder dysfunction, many have not been adequately studied in the HNC population. Therefore, the appropriateness of test score interpretation is unknown. The DASH, QuickDASH, NDII, SPADI, and the UW-QoL (shoulder subscale) have been suggested as appropriate PROs to quantify shoulder dysfunction in this patient population.[2] These recommendations are based upon expert opinion and sample-specific psychometric properties derived from CTT methodology, which in most cases, are not generalizable to the HNC population. Recent studies utilizing Rasch analysis suggest several limitations in scale item hierarchy, response scale structure, and item fit for the DASH, QuickDASH, NDII, and the SPADI in patients with HNC.[19,21] Similar findings have been reported for the DASH, QuickDASH and SPADI in other patient populations.[10,11,13-15,20] These limitations suggest weakness of the construct validity and precision of the scales. Further research is, therefore, needed to validate the Academy of Oncologic Physical Therapy EDGE Task Force's recommendations for PROs to quantify shoulder function in patients following neck dissection surgery for HNC.

This research study seeks to add to the construct validity of the recommended PROs by demonstrating that each test's items adequately measure the latent construct of interest - shoulder function/dysfunction. Person and item measures will provide information related to how well the tests target the population of individuals who have received neck dissection surgery for HNC. They will also provide information related to floor and/or ceiling effects, gaps in measuring individual ability levels, and redundant test items that increase responder burden. Fit statistics will allow us to determine how each item contributes to the latent construct, and whether the assumptions of unidimensionality and item interdependence are violated. Unidimensionality will be further assessed through PCA, item fit and DIF. Scale reliability will be reported through person and item reliability estimates.

The results of this study will further the work of the APTA's Academy of Oncologic Physical Therapy Head & Neck EDGE Task Force[2] by increasing the ability of clinicians and researchers to interpret an individual's disability level, in addition to the PRO's ability to measure change. While assessment of scale responsiveness is outside the scope of this research study, findings within our analysis of large floor or ceiling effects could negate the need to perform further studies, allowing focus on more relevant measures to the HNC patient population.

Chapter 3: Methodology

Introduction to the Chapter

Chapter 3 describes the methodology for this research study. The research design including a description of the sample recruiting and the study personnel, research setting, recruitment methods, enrollment process including informed consent, data collection, and data analysis will be discussed. Formats for presenting results will be described, in addition to resources used for this study.

Research Method

This research study was a multi-site, cross sectional, questionnaire-based psychometric study. Participants were recruited using convenience sampling from a population of patients with a history of neck dissection procedures for management of head and neck cancer (HNC).

Specific Procedures Employed

Ethical Approval and Study Registration

Prior approval from the Institutional Review Boards (IRB) at the Mayo Clinic and Nova Southeastern University was obtained (Appendix 1). The Mayo Clinic IRB served as the IRB of Record (IRB number 15-005266). This research was presented at the Neurology Discipline Oriented Group Committee at the Mayo Clinic on August 25, 2015, which oversees research studies generated from the Neurology and Physical Medicine and Rehabilitation departments at the Mayo Clinic that involve the Clinical Studies Unit. The study is listed on ClinicalTrials.gov (NCT02554968).

Participants

To be eligible for this study, subjects must have had either a unilateral or bilateral neck dissection procedure for management of HNC within the past two weeks to 18 months, and endorse some level of shoulder impairment (i.e. Answer "yes" when asked "are you currently experiencing any shoulder weakness, stiffness or discomfort as a result of your neck surgery?"). Eligible subjects were between 18-90 years of age and fluent in the English language allowing for completion of the study-related forms and questionnaires. Subjects were excluded if they denied shoulder impairment as a result of the recent neck dissection surgery (i.e. answered "no" to the previously stated screening question). Rasch analysis requires a heterogeneous sample and range of ability levels related to shoulder function, therefore subjects with a severed Spinal Accessory Nerve (SAN) and those who were current receiving or who had already received radiotherapy, chemotherapy or combined modality treatments were not excluded. In addition, subjects with any stage or type of HNC were eligible, as long as a selective neck dissection (SND), modified radical neck dissection (MRND), or radical neck dissection (RND) procedure was included as part of the medical management.

Recruitment

One hundred and eighty-two subjects were enrolled in this study. According to Linacre, 108 – 243 subjects are required to gain a 99% confidence that the estimated item difficulty is within ±½ logit of its stable value.[352] Pilot data of 131 questionnaires demonstrates poor response option utilization by HNC survivors for Disability of the Arm, Shoulder and Hand (DASH) and QuickDASH test items which were more difficult to endorse (response options 4 and 5),[19] therefore we aimed to recruit a larger sample in an attempt to gain a more heterogeneous sample related to ability level, and to account for incomplete questionnaires. Attrition was not

considered as a factor in recruitment planning for this study because subjects were only asked to complete the questionnaires on one occasion with no scheduled follow-up required.

Subject recruitment was completed using two methods: (1) in-office recruitment from the Otolaryngology, Radiation Oncology, and Physical Medicine and Rehabilitation departments at the Mayo Clinic in Phoenix, Arizona, Jacksonville, Florida, and Rochester, Minnesota; and (2) mailed questionnaires. We first attempted to complete enrollment through in-office recruitment, however slow recruitment prompted the addition of mailed questionnaires. Goldstein and colleagues report an 80% response rate in a questionnaire-based study of individuals with HNC.[353] During in-office recruitment, potential subjects were screened for eligibility by designated medical providers at the time of their medical appointment. If deemed eligible, the potential subject was provided with an Oral Consent Template (Appendix 2), Health Insurance Portability and Accountability Act (HIPAA) Authorization to Use and Disclose Protected Health Information form[a] (Appendix 3), and questionnaires to complete. In an effort to prevent enrollment of the same subject twice, the Oral Consent Template asks the participants to return the uncompleted questionnaires if they have already completed them at another appointment. The second method of recruitment utilized quarterly electronic reports generated by the Mayo Clinic Health Sciences Research Department, which included any patient billed for a neck dissection procedure within the Mayo Clinic Enterprise. The report did not include individuals outside of the eligible age range, or those with legal, financial or deceased administrative flags. This report was screened by the Primary Investigator to confirm accuracy and remove

[a] Communication with the Mayo Clinic IRB has determined that a full-length consent form was not necessary for this minimal risk study. Study personnel provided a brief statement such as: "Your input about your shoulder pain or weakness is important to us. We would appreciate if you would complete these questionnaires." The study personnel would then offer the subjects the oral consent template, HIPAA release form, and questionnaires to complete.

individuals receiving neck dissection surgery for an indication other than HNC. Packets were subsequently mailed to the remainder of individuals on the list. Packets included the following information: informational letter (Appendix 4) modified from the Oral Consent Template, the HIPAA form, and questionnaires. A self-addressed, postage-paid envelope was also included. The order of forms within the packet was the same for each subject enrolled through in-office or mailed recruitment.

Data Collection

For in-office recruitment, after providing consent subjects were asked to complete a demographic questionnaire, which captured the following information: self-reported height and weight, ethnicity, hand dominance, and previous and/or current receipt of treatment for shoulder dysfunction (Appendix 5). Additional information was gleaned from the electronic health record (EHR) by the Primary Investigator to further describe the sample including the subject's age and gender, state of residence, location of treatment (Arizona, Florida, or Minnesota), date of surgery, tumor type and stage, surgery received (including bilateral/unilateral neck dissection, levels dissected, and status of the SAN), and additional treatments received (including radiotherapy, chemotherapy, or combined treatments) (Appendix 6). Subjects then completed the four shoulder-related patient-reported outcome measures (PRO): DASH, Shoulder Pain and Disability Index (SPADI), Neck Dissection Impairment Index (NDII), and the University of Washington Quality of Life (UW-QoL) (shoulder subscale). The single item UW-QoL (shoulder subscale) was placed on the demographics questionnaire document in an effort to decrease responder burden and the number of forms to complete (Appendix 5). Test items for the QuickDASH were obtained directly from the DASH, and therefore were not completed by the subject. Derivation of QuickDASH scores from DASH scores has been described previously in

the literature.[10,245] All questionnaires were collected and sent to data coordinators who were responsible for uploading and storing data in REDcap (Research Electronic Data Capture) hosted at the Mayo Clinic.[354] Upon completion of data collection, a data coordinator completed a random quality check for data entry for 10% of subjects.

This research study utilized two data coordinators. The first was employed as an Associate Clinical Research Coordinator by the Mayo Clinic Clinical Studies Unit. Her time was funded by grant money described later. The second was a rehabilitation technician in the Department of Physical Medicine and Rehabilitation.

Outcome Measures

The DASH, QuickDASH, SPADI, NDII and the UW-QoL (shoulder subscale) were utilized for this study. Each of these were described in detail previously, and therefore are not described here. For the purposes of this study, there are a few considerations to mention. The subjects in this study were not asked to complete the two optional work and performing arts modules for the DASH and QuickDASH. The NDII rating scale is inverted when compared to the other PROs in this study. The test score with the regular rating scale was used for Classical Test Theory (CTT) comparisons, however individual test item responses were transformed to be consistent with the other measures for Rasch analysis. Techniques to manage missing data are not reported for the NDII, therefore the same requirement for the 11-item QuickDASH of a 90% response rate was required (nine items) for calculating the test score.

Data Analysis

Descriptive statistics were calculated using JMP® Pro 13.0.0 (Copyright © 2016 SAS Institute Inc., Cary, North Carolina) to report sample characteristics and mean population test score values. A correlational analysis was also utilized to compare the relationships between

each of the PROs and the single-item UW-QoL (shoulder subscale). Prior to Rasch analysis, NDII item responses were transformed to coincide with the response structures of the other measures. A high score on the DASH, QuickDASH, SPADI and the UW-QoL (shoulder subscale) indicate greater disability, whereas a high score on the NDII suggests lower disability. Rasch analysis will be performed using Winsteps Rasch Measurement computer program, version 3.81.0 (Winsteps, Beaverton, Oregon, USA) to assess the reliability, construct validity, and overall appropriateness of test score interpretation of the DASH, QuickDASH, SPADI, and the NDII in patients experiencing shoulder dysfunction following neck dissection surgery for HNC. Specifically, we assessed scale dimensionality, through consideration of principal components analysis (PCA), item and person fit, and Differential Item Functioning (DIF), response scale structure, and item and person separation and reliability. The analysis was guided by the previously mentioned investigational questions.

Scale Dimensionality

Dimensionality of the scales was first assessed using PCA, then though consideration of item fit and DIF. PCA standardized residuals with an eigenvalue of 2.0 or higher in the first contrast triggered further assessment of item loading coefficients. Specifically, groups of items with a loading coefficient of ≥ 0.40 were compared to the model for the presence of additional dimensions. Exploratory analysis of PCA, item fit and summary statistics was subsequently utilized to consider whether subscales should be utilized for further analysis.

Further consideration of scale dimensionality included analysis of item and person fit and DIF. Person or item misfit was determined by analysis of person or item infit and outfit statistics (Mean-Square , MNSQ; z-standard, ZSTD). For this study a MNSQ range of 0.6 to 1.4 was deemed acceptable. When the MNSQ fell outside of this range, the ZSTD statistic was analyzed.

When the ZSTD was > ±2, person or item misfit was present.[177] A qualitative assessment of misfitting items was then performed. Further exploration of misfitting items occurred through analysis of PCA, item fit and summary statistics with misfitting items eliminated to determine impact of overall scale functionality. Person misfit was also considered to determine whether there is a relevant cause or reason that the person answered test items differently than expected.

The presence of DIF may also suggest the presence of additional constructs or dimensions in a measure. DIF by age, which was arbitrarily set at 18-64 years and 65-85 years, and gender (male and female) was considered in this analysis. For the purposes of this study, items with a DIF contrast (effect size) of >0.64 ($p < 0.05$) were considered for potential DIF.[180] Positive findings of items with DIF were qualitatively considered based upon the recommendations of Linacre.[181]

Response Scale Structure

To assess response scale structure for the DASH, QuickDASH, SPADI, and NDII, this study considered response category utilization, distribution of responses across options, scale calibration, average measure values and item fit through the category structure output provided by Winsteps.[8,183] Step calibration and average measures should sequentially increase or decrease with item difficulty for each response option. A response category outfit value of greater than 2.0 suggests that there is more unexplained variance than explained variance in a category, and therefore limits the accuracy, stability and interpretability of the measure.[182]

In the case of unequal utilization of response categories, the presence of disordered response options and excess response option variance, this study considered the appropriateness of collapsing the scale structure to improve the precision of the measure and decrease responder burden. Subsequently, Rasch analysis, including item and person separation and reliability,

person and item fit, scale hierarchy, and scale dimensionality, were run using the optimized response scale structure. If collapsing categories failed to improve scale reliability and separation indices, analysis proceeded using the original response scale structure.

Scale Hierarchy

Rasch half-point threshold maps were analyzed to describe the scale hierarchy of the PROs. For the purposes of this study, a floor or ceiling effect occurred when 5% of the sample scored the lowest or highest possible score.[355] The presence of gaps or item redundancies are reported for each measure. The easiest and most difficult items for each PRO for this sample are also reported based upon logit (standard error, SE) measures.

Item and Person Separation and Reliability

The 'REAL' estimate was utilized for this study when reporting reliability estimates. In Winsteps, the 'REAL' estimate provides the person reliability coefficient with extreme scores removed. The 'MODEL' estimate does not remove extreme scores, therefore Boone and colleagues suggest using the REAL estimate when reporting.[185] A person separation of greater than 2 with a person reliability of greater than 0.8, and an item separation of greater than 3 with an item reliability of greater than 0.9 were utilized as a cutoff for this study.[186]

Formats for Presenting Results

Description of the Sample

Sample demographics, including mean, standard deviation (SD) and range for continuous data and frequencies for categorical data, are reported for the following characteristics: age, gender, state of residence, ethnicity, body mass index (BMI), tumor type and stage, time since surgery, surgery characteristics (unilateral or bilateral neck dissection, levels dissected, status of

SAN), and additional interventions received (radiotherapy, chemotherapy, chemoradiotherapy, physical and/or occupational therapy).

Description of PRO Test Score Results

A description of the number of questionnaires completed, the number of incomplete questionnaires subsequently removed from analysis, and the number of questionnaires analyzed using Rasch are provided. Descriptive statistics including mean, SD, and range for the DASH, QuickDASH, NDII, SPADI and UW-QoL (subscale) test scores are reported. Results of a correlational study of the five measures are reported.

Examination of the Research Question

The results of the Rasch analysis, related to scale dimensionality, response scale structure, and reliability are presented for each PRO to answer the research questions. A combination of text, tables and figures are utilized. Information regarding response category utilization and item fit are provided in table format; and a description of scale hierarchy including gaps, redundancies and floor/ceiling effects are shown using figures of the half-point threshold maps for each PRO.

Examination of Additional Research Questions

A secondary aim of this research study was to either confirm or modify the recommendations made by the Academy of Oncologic Physical Therapy Head & Neck Evaluation Database to Guide Effectiveness (EDGE) Task Force for shoulder-related PROs best utilized in the HNC population.[2] If the preliminary analysis of the five recommended PROs fails to provide a single PRO that can be used appropriately, secondary analysis may continue with the aim to provide a recommendation for a combination of PROs that can adequately quantify shoulder dysfunction in this patient population.

Dissemination of Results

Pilot work for the dissertation was presented at the American Physical Therapy Association's (APTA) Combined Section Meeting in Indianapolis, Indiana in 2015,[19] and initial analysis of the DASH, QuickDASH, SPADI and NDII of this work (sample of 90 subjects) was presented at the APTA's Combined Section's Meeting in San Antonio, Texas in 2017.[356] A full-length manuscript detailing the pilot work presented in 2015 is under peer-review. Upon completion of the dissertation and defense, it is anticipated that 4-5 additional manuscripts will be submitted for publication, which detail the results of Rasch analysis for the DASH, QuickDASH, SPADI, and NDII, as well as the usability of the UW-QoL (shoulder subscale) as a screening tool An additional manuscript suggesting a combination of measures for quantification of shoulder-related dysfunction in the HNC population will also be considered. Educational sessions, platform and poster presentations at national and international meetings will also be considered.

Resources Used

Grant support of $5000 was awarded through the APTA's Academy of Oncologic Physical Therapy (Appendix 7). This funding helped to cover costs associated with mailed questionnaires and research coordinator time through the Mayo Clinic Clinical Studies Unit. The Mayo Clinic Department of Physical Medicine and Rehabilitation supported costs once the grant award was utilized, including fees for reports, generated quarterly, from which potential subjects were recruited, cost of mailed questionnaires, and personnel support for data entry. Resources were not allocated for the time it took the Primary Investigator to perform clerical tasks such as putting together and mailing questionnaires, or data mining of the EHR. Data coordinators were responsible for inputting data into the computer-based data management system (REDcap).

Other required resources included software and supplies. Software requirements for this study included Winsteps, REDcap, and JMP. Winsteps is provided at no cost by Nova Southeastern University, and REDcap and JMP are provided at no cost by the Mayo Clinic.

Chapter 4: Results

Introduction to the Chapter

Chapter 4 details the results to the study. The individual research questions and study aims are addressed. Tables and figures are utilized to help organize the chapter.

Description of the Sample

Sample Size

Subjects were recruited between September 2015 and August 2017. Fifty subjects were enrolled through in-office recruitment. A total of 721 questionnaire packets were mailed and 229 were returned (31% response rate). Of the 229 packets returned, 75 declined participation, 19 individuals signed the Health Insurance Portability and Accountability Act (HIPAA, Appendix 3) form but denied shoulder pain and therefore were considered screen failures, one individual with shoulder impairment completed and returned the packet but the date of signature was outside of the eligibility timeframe of 18 months and was also considered a screen failure. Two returned packets were notifications of patient expiration. One hundred and thirty-two individuals were accrued through a mailed questionnaire. A total of 182 individuals were accrued for this study. Informed consent was implied through completion of the questionnaires and signing of the HIPAA form (Figure 1).

Of the 182 returned packets, three packets were incomplete. Two packets did not include completed demographic information. On each PRO, participants skipped test items, and in very rare cases skipped the entire questionnaire [Disability of the Arm, Shoulder and Hand (DASH)/QuickDASH, N=2; Shoulder Pain and Disability Index (SPADI), N=4; Neck Dissection Impairment Index (NDII=1)]. The University of Washington Quality of Life (UW-QoL) (shoulder subscale) was skipped by five participants. See Table 2 for details regarding skipped

test items and skipped questionnaires. Because Rasch analysis focuses on independent items within a test rather than a test summary score, incomplete packets were retained for data analysis. However, when reporting mean summary test scores for each of the PROs questionnaires with missing items were eliminated consistent with scoring requirements.

Figure 1. Research Study Recruitment

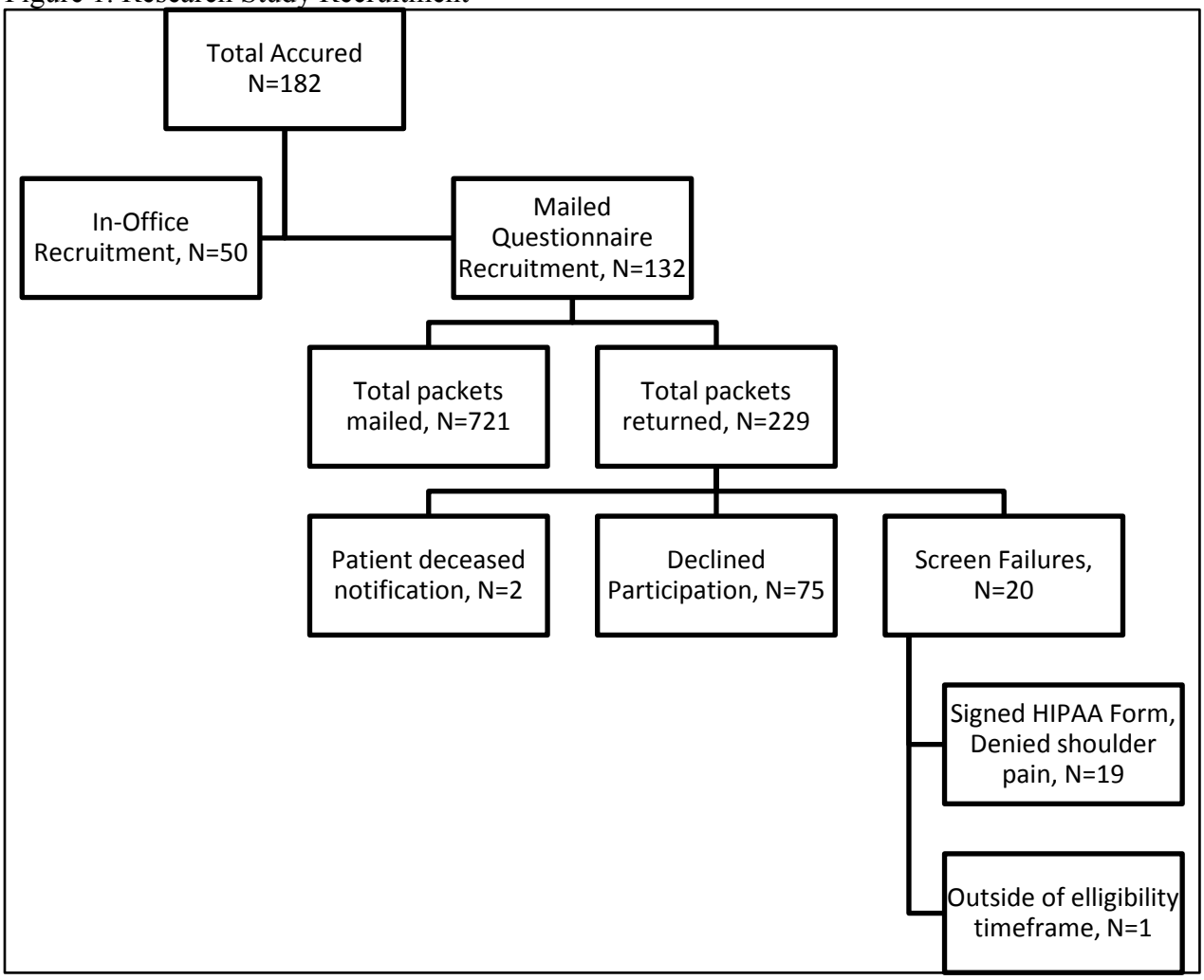


Table 2. Description of Incomplete Questionnaires and Skipped Items

	# of incomplete questionnaires	# of skipped items	# of questionnaires eliminated from analysis
DASH	31	1: 12 2: 8 3: 2 *4: 2* *5:1* *6: 2* *11: 1* *12: 1* *All: 2*	9 (27 items required for calculation of summary score)
QuickDASH	18	1: 12 *2: 3* *4: 1* *All: 2*	6 (10 items required for calculation of summary score)
SPADI-pain	5	1: 1 *All: 4*	4 (no guideline based upon number needed to score for SPADI)
SPADI-disability	9	1: 3 2: 1 3: 1 All: 4	4
SPADI-total	9	1:2 2:2 3:1 *All: 4*	4
NDII	7	1: 5 2: 1 *All: 2*	3
UW-QOL (shoulder subscale)	6	N/A	0

*Italics indicate number of test items missing, that result in removal from analysis for test score reporting.

Demographics

Table 3. Sample Demographics

Age at enrollment (mean, SD, range)		62.67	(11.47, 33-87)
Gender (N,%)	Male Female	143 39	79% 21%
Ethnicity (N, %)	African American Asian American White, Non-Hispanic White, Hispanic	3 2 161 13	2% 1% 88% 7%

	Middle Eastern Other	0 3	0% 2%
Height – cm (mean, SD, range)		175.48 cm	(10.84, 97-198.12)
Weight – kg (mean, SD, range)		81.18 kg	(16.13, 44-145)
BMI – kg/m^2 (mean, SD, range)		26.38 kg/m^2	(5.02, 17.63-53.26)
Hand dominance* (N, %)	Right Left Ambidextrous	156 14 7	88% 8% 4%
Days since surgery (mean, SD, range)		174.07	(138.43, 14-567)

Abbreviations: standard deviation (SD), number (N), centimeters (cm), kilograms (kg), body mass index (BMI), meters (m)
*N=177 (5 did not report hand dominance)

Twenty-five states within the United States were represented in the sample. The majority of individuals enrolled held primary residency in Arizona (N = 58) or Minnesota (N = 52). The remainder of the sample was fairly representative of the United States regions, with the exception of the Northeast, with 23 participants from the South [Florida (13), Georgia (6), Tennessee (1), Texas (1), Arkansas (1), Mississippi (1)], 36 additional participants from the Midwest [Iowa (16), Wisconsin (5), Michigan (5), North Dakota (4), Illinois (3), Indiana (1), Nebraska (1), and South Dakota (1)], and 11 additional participants from the West [Montana (3), New Mexico (2), Nevada (2), California (1), Colorado (1), Idaho (1) and Wyoming (1)]. Hawaii and Alaska were represented with one participant each. Consistent with the greatest representation being from the Midwest states, the majority of cancer care occurred at Mayo Clinic in Rochester, Minnesota (N = 98, 54%). Sixty-six individuals (36%) received their cancer treatment at Mayo Clinic in Phoenix, Arizona and 17 individuals (9%) received care at Mayo Clinic in Jacksonville, Florida.

Cancer Diagnosis

Any individual receiving a neck dissection procedure was eligible for this study. The majority of individuals enrolled were diagnosed with cancer of the oropharynx or hypopharynx (N = 85). Cancers of the oral cavity and larynx (N = 49), parotid (N = 17), nasopharynx (N = 1), nasal cavity and paranasal sinuses (N = 2) and thyroid (N = 6) were included. Some skin cancers of the head were also included (N = 11). Eleven participants receiving neck dissection were characterized as having a cancer diagnosis of "other." Details regarding levels dissected are available in Table 4 and tumor staging are available in Table 5. Of note, review of the electronic health record (EHR) in many cases revealed incomplete staging information. In many cases, lack of the TNM classification was for individuals with diagnoses staged differently than the cancers of the oropharynx, hypopharynx, oral cavity and larynx. In only 14 cases was the Spinal Accessory Nerve (SAN) sacrificed intraoperatively (right, N=5; left, N= 9).

Table 4. Description of Surgery in the Study Population

	Right Neck Dissection	Left Neck Dissection	Bilateral Neck Dissection
Number of surgeries performed	83	58	41
Levels dissected	N = 124	N = 99	
Level 1A/1B only	2	1	
Level 2A/2B only	2	2	
Level 1 & 2	4	1	
Levels 1 – 3	10	6	
Levels 1 – 4	28	27	
Levels 1 – 5	9	9	
Levels 2 & 3	3	7	
Levels 2 - 4	51	34	
Levels 2 – 5	11	10	
Levels 3 – 5	1	1	
Levels 4 – 5	0	0	
Other	3	1	

Table 5. Classification of Cancer for Study Population

TNM classification		Number of Participants with Classification	Percent of Sample
T (N=155)	TX	5	3%
	T0	0	0%
	Tis	0	0%
	T1	47	30%
	T2	53	34%
	T3	21	13%
	T4	29	19%
N (N=151)	NX	3	2%
	N0	32	21%
	N1	20	13%
	N2	86	57%
	N3	10	7%
M (N=161)	MX	5	3%
	M0	156	97%
	M1	0	0%

Cancer Treatment

Sixty-eight individuals received surgery as a single modality intervention for their head and neck cancer (HNC) treatment, whereas 52 individuals received adjuvant radiotherapy and 61 individuals received combined modality radiation and chemotherapy. Adjuvant interventions for one individual are unknown. Although adjuvant radiotherapy was recommended for this individual, review of the EHR was unclear as to whether the intervention was actually pursued outside of the Mayo Clinic Enterprise.

Endorsing shoulder pain, weakness or stiffness was a requirement for eligibility in this study. Ninety-five participants (52.2%) reported right shoulder impairment, 70 (38.5%) reported left shoulder impairment and 17 individuals (9.3%) reported bilateral shoulder impairment. The majority did have their shoulder impairment addressed by a care provider through intervention or exercise prescription. Fifty-eight percent (N=105) reported receipt of care from a physical therapist. Physicians (13%), massage therapists (8%), nurses (7%), occupational therapists (7%),

chiropractors (7%), acupuncturists (1%) and personal trainers (1%) also participated in the management of shoulder impairment. Only 53 study participants (29%) reported receiving no interventions related to their shoulder pain, stiffness, or weakness.

Data Analysis

Classical Test Theory

Descriptive Statistics of PROs

Classical Test Theory (CTT) methodology was utilized to report test score average (standard deviation, SD), standard error of the measure (SEM), and 95% confidence interval (CI) for each of the measures (Table 6). These statistics were generated after removing questionnaires that did not meet the minimum requirement for answered items for scoring, as indicated in Table 2. Frequency of category utilization for the DASH, QuickDASH, SPADI and NDII will be reported later. Distribution of responses across response option for the UW-QoL (shoulder subscale) are outlined in Table 7.

Table 6. PRO Average Test Scores, SEM and 95% CI using CTT

	DASH	Quick-DASH	SPADI-Pain	SPADI-Disability	SPADI-Total	NDII	UW-QOL shoulder
Mean	28.83	28.84	35.98	26.61	30.05	57.58	47.67 (mean) 70 (median, mode)
SD	18.00	19.06	25.48	22.32	22.59	22.54	25.47
SEM	1.37	1.44	1.91	1.67	1.69	1.68	1.92
95% CI	26.13, 31.53	26.00, 31.67	32.21, 39.75	23.31, 29.91	26.71, 33.39	54.27, 60.90	43.88, 51.46
Sample size	173	176	178	178	178	179	176

Table 7. Distribution of Response for the UW-QoL (shoulder subscale)

Shoulder (Item)	**Scoring Value (points)**	**Frequency of Use (N)**
1. I have no problem with my shoulder.	100	3
2. My shoulder is stiff but it has not affected by activity or strength.	70	86
3. Pain or weakness in my shoulder has caused me to change my work/hobbies.	30	69
4. I cannot work or do my hobbies due to problems with my shoulder.	0	18

Correlational Analysis

Correlational analysis of the five PROs was completed (Table 8). As expected, the DASH and QuickDASH were found to have a statistically significant, strongly positive relationship. Similarly, the SPADI-total score and the disability and pain subscores were found to have a strong positive correlation. A significant, but weak to moderate negative correlation was found for the NDII and the DASH, the NDII and the three SPADI scores, and the UW-QoL and the three SPADI scores.

Table 8. Correlational Analysis of the PROs

	DASH	QuickDASH	UWQOL	SPADI-Disability	SPADI-pain	SPADI-total	NDII
DASH							
QuickDASH	0.98***						
UWQOL	-0.0067	0.004					
SPADI-disability	0.11	0.12	-0.26***				
SPADI-pain	0.09	0.099	-0.19**	0.88***			
SPADI-total	0.10	0.11	-0.24**	0.96***	0.98***		
NDII	-0.15*	-0.14	0.054	-0.19*	-0.17*	-0.18*	

*p<0.05; **p<0.01; ***p<0.001

Because there was no correlation between the UW-QoL (shoulder subscale) and the DASH, QuickDASH and NDII, further analysis to determine the usability of the subscale as a screening tool, or predictor of shoulder impairment, was not pursued. The usability of the UW-QoL (shoulder subscale) based upon its weak to moderate correlation with SPADI was then

considered. Because of the unequal distribution of category 1 and 4 utilization, it was decided that further analysis of the UW-QoL (shoulder subscale) as a screening tool was not appropriate. This sample does not provide a clear representation of shoulder impairment across ability levels.

Rasch Analysis

Disabilities of the Arm, Shoulder and Hand

See Appendix 8 for DASH and QuickDASH data including: test items, item stems, item logit values, item fit statistics, and response category utilization.

Scale Dimensionality

Principal component analysis (PCA) of the DASH showed that 60.9% of the variance in the scores was explained by the measure with an eigenvalue of 3.2, suggesting the presence of greater than one construct measured by the patient-reported outcome measure (PRO). Assessment of loading coefficients for the items meeting the criteria for further assessment (>0.40) revealed 5 items which loaded in the positive direction and 4 items which loaded in the negative direction (See Table 9). Comparison of these positive- and negative-loading items reveals 2 distinct subdomains: functional activities and symptoms. Subsequent analysis of scale dimensionality using PCA was completed on a 24-item function subscale (Items 1-23, 30) and a 6-item symptom subscale (Items 24-29). PCA of the 6-item symptom subscale showed an eigenvalue of 1.6; however the 24-item function subscale showed an eigenvalue of 2.9. Five items loaded in the positive direction and one item in the negative direction.

Table 9. Assessment of Dimensionality Using Principal Component Analysis of the DASH

Item	**Loading**
Item 07, *heavy chores*	0.58
Item 06, *object overhead*	0.57
Item 12, *change bulb*	0.57
Item 15, *don sweater*	0.45
Item 13, *style hair*	0.40
Item 24, *pain*	-0.56
Item 26, *tingling*	-0.56
Item 25, *pain with activity*	-0.46
Item 29, *sleep*	-0.44

To further examine the observed patterns of item loadings, exploratory factor analysis was completed using JMP® Pro 13.0.0 (Copyright © 2016 SAS Institute Inc., Cary, North Carolina). This did not show a clear pattern; therefore additional PCA was run based upon previously published work suggesting subscales of the DASH: manual functioning, disability secondary to limitation of motion, and symptoms.[14] While the PCA for each of the three scales fell below an eigenvalue of 2.0, summary statistics for each were poor compared to the full test and analysis of item fit issues for each of the scales increased.

Five test items were found to misfit the model for item fit on the 30-item DASH (Appendix 8). Items 1: *open jar,* and 21: *sexual activities* were found to have problems with infit and outfit, whereas Items 4: *prepare a meal*, 26: *tingling* and 29: *sleep* only misfit the model with outfit statistics. Person misfit was substantial for this sample. Nineteen percent (n = 35) misfit the model with high (n=17) or low (n=18) infit statistics; and 16% (n = 28) misfit the model with high (n=13) or low (n=15) outfit statistics. Based upon these findings, and consistent with analysis of pilot data currently under peer-review (Eden, Kunze, Cheng, unpublished data, 2018) it was determined that the DASH is not a suitable scale for the current sample of the HNC population and further analysis of the DASH or potential subscales did not occur.

Scale Structure

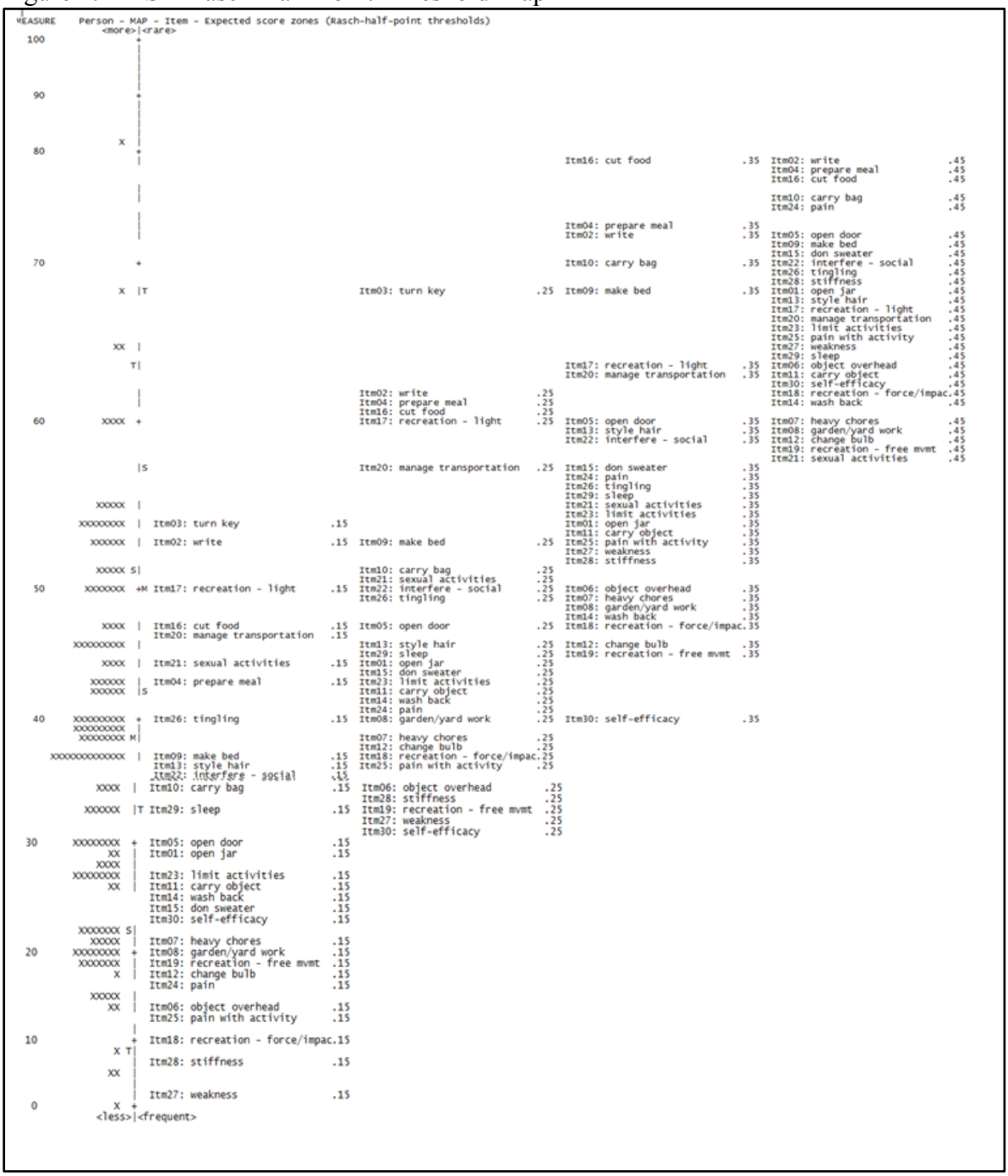
To further explain the proposed argument that the DASH is not suitable for this sample of the population, a description of the 30-item DASH scale structure is provided here. Consideration of the Rasch half-point threshold map (Figure 2) indicates that the DASH is able to measure across the majority of ability levels, except for at the highest disability levels. No floor or ceiling effect was identified. Nine gaps in the scale structure and 26 areas of redundancy were also identified.

In this population, Item 2: *write* and Item 16: *cut food* were the easiest items (hardest to endorse as being difficult) with a logit (standard error) value of 66.6 (1.5) and 66.5 (1.4) respectively. Item 18: *recreation-force/impact* (logit 38.7, SE 1.0) and Item 27: *weakness* (38.7, 1.1) were the most difficult items. The DASH did not meet criteria for an optimal rating scale. Twenty-one (70%) of test items did not meet minimum criteria for response option utilization of 10 responses each. Additionally, no test item was found to have uniformity of distribution across response options. Forty-three percent of test items were found to have either disordered step calibrations or average measures, or both. Five items were found to have a response category outfit mean square statistic (MNSQ) value exceeding 2.0 (Items 1: *open jar*, 2: *write*, 21: *sexual activities*, 26: *tingling*, and 29: *sleep*).

Reliability

The person separation and reliability index for the DASH (4.26, $\alpha = 0.95$) are large enough to adequately classify individuals into ability levels. The item separation and reliability index (7.30, $\alpha = 0.98$) is large enough to confirm scale hierarchy. Estimates for reliability for person separation in Winsteps is excellent ($\alpha = 0.96$).

Figure 2. DASH Rasch Half-Point Threshold Map



QuickDASH

Scale Dimensionality

PCA of the QuickDASH showed that 58.9% of the raw variance was explained by the measure, with an eigenvalue of 2.0. An eigenvalue of <2.0 is an acceptable cut-off for PCA, therefore further analysis was not performed.

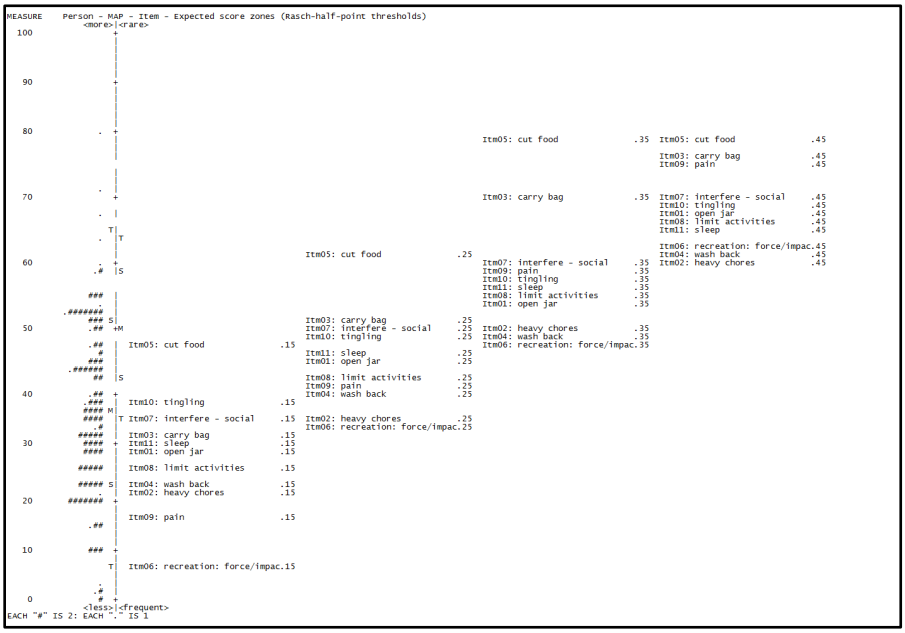
Dimensionality of the QuickDASH was further analyzed through consideration of item and person fit statistics and Differential Item Functioning (DIF) (age and gender). The scale was found to have two misfitting items (Appendix 8). Item 1: *open jar* misfit the model for infit (MNSQ 1.42, z-standard, zstd 3.4) and outfit (1.49, 3.5), and Item 10: *tingling* misfit only for the outfit statistic (1.47, 2.2). Analysis of DIF for gender found issues on three items. Females answered Items 1: *open jar* and 3: *carry bag* as being more difficult than males, and males answered Item 4: *wash back* as more difficult than females. No items were found to be significant for DIF based upon age. Based upon concerns with item misfit and DIF by gender for Item 1: *open jar*, analysis of the test with removal of this item was performed. The eigenvalue for PCA remained at 2.0, and no significant changes in summary statistics were identified. Further analysis of item fit in the 10-item scale revealed item misfit (infit and outfit) for Item 10: *tingling*. DIF for gender with removal of Item 1: *open jar* continued for Item 3: *carry bag* but was no longer a concern for Item 4: *wash back*. At this point, it was decided to retain all 11 items for further analysis and reporting for the QuickDASH.

Person misfit was identified in 16 cases (approximately 9%), 15 individuals (8%) misfit the model for infit (11 with high infit and four with low infit statistics) and nine individuals (5%) for outfit (seven with high infit and two with low infit statistics).

Scale Structure

Analysis of the scale structure for the QuickDASH was completed through observation of the Rasch-half point threshold map (Figure 3), which demonstrates better ability of the scale to classify individuals with low to moderate levels of disability.

Figure 3. QuickDASH Rasch Half-point Threshold Map



Results showed that Item 5: *cut food* was the easiest item (least likely to be endorsed as being difficult) with a logit (standard error) value of 66.5 (1.5), and Item 6: *recreation–force/impact* was the hardest item (most likely to be endorsed as being difficult) with a logit value of 37.8 (1.0). A floor and ceiling effect were not present. The Rasch-half point threshold map revealed 10 gaps in the scale structure, and six occurrences in which more than one test item response option measures the same ability level (redundancies).

Analysis of category utilization of the QuickDASH identified eight of the 11 items had fewer than 10 responses in at least one of the response option categories. Items 1 and 7-11 did not achieve the required 10 responses in the fifth response option category, and Items 3: *carry bag* and 5: *cut food* lacked 10 responses in the fourth and fifth category. In fact, in Item 5: *cut*

food, response option 4 was not selected by any respondent. It was also noted that all items lacked uniform distribution of response category utilization; although the average measures did advance monotonically by category for each of the 11 test items. There were two occurrences in which disordered step calibration was identified, and one occurrence in which response category outfit MNSQ values exceed 2.0 (Item 1: *open jar*) (Appendix 8).

Collapsing response options 4 and 5 improved category utilization to meet the 10-response requirement for all items except for Items 3: *carry bag* and 5: *cut food*. One of the 11 test items, Item 2: *heavy chores*, gained uniformity in utilization of response options (range 36 to 50 responses per option). With collapsing response options 4 and 5, Item 3: *carry bag* was found to have disordered average measures; however prior occurrences of disordered step calibration and high outfit MNSQ values were alleviated. Person separation and reliability statistics did not change with collapsing response options (collapsed: 2.42, $\alpha = 0.52$ versus no change: 2.49, $\alpha = 0.86$). The item separation index decreased to 6.19 ($\alpha = 0.97$), keeping it well above the cut-off for an acceptable scale. PCA and item fit did not change with this modification.

Reliability

The person separation and reliability index for the QuickDASH (2.49, $\alpha = 0.86$) are large enough to adequately classify individuals into ability levels. The item separation and reliability index (6.60, $\alpha = 0.98$) is large enough to confirm scale hierarchy. Estimates for reliability for person separation in Winsteps is excellent ($\alpha = 0.90$).

Shoulder Pain and Disability Index

See Appendix 9 for SPADI data including: test items, item stems, item logit values, item fit statistics, and response category utilization.

Scale Dimensionality

PCA of the SPADI was first run combining the pain and disability subscales. The measure explained 75.2% of the raw variance (eigenvalue 2.3). Two items were found to load in the positive direction, and four items were found to load in the negative direction (Table 10). PCA was subsequently run on the subscales separately. The measure explains 78.7% of the raw variance and yielded an eigenvalue of 1.8 for the first contrast of the pain subscale. For the disability subscale, 74.8% of the raw variance was explained by the measure, with an unexplained variance for the first contrast of an eigenvalue of 2.2.

Table 10. Principal Components Analysis of SPADI (Pain and Disability)

Item	**Loading**
Item 03: *pain-shelf*	0.65
Item 01: *pain-worst*	0.58
Item 10: *disability-don pants*	-0.64
Item 8: *disability-don shirt*	-0.53
Item 9: *disability-buttons*	-0.52
Item 13: *disability-back pocket*	-0.42

Dimensionality of the SPADI was further analyzed through item and person fit. For the full scale, Item 12: *disability-heavy object*, misfit the model with high infit (MNSQ 1.51, zstd 3.7) and outfit (1.68, 4.4) statistics (Appendix 9). Removal of Item 12: *disability-heavy object* did not impact PCA findings or significantly impact summary statistics. The item was therefore retained for further analysis. When analyzing the SPADI-disability subscale, Item 12: *disability-heavy object* was, again, found to misfit the model with high item infit (1.55, 4.0) and outfit (1.63, 4.3) statistics. No items misfit the SPADI-pain subscale model (Appendix 9). DIF for gender was found for Item 6: *disability-wash hair*, where females found the item to be more difficult than males. No items were found to have DIF for age.

Person fit for the full SPADI scale found misfit for infit for 19% of the sample (n=33) and outfit for 15% (n=27). For infit, 14 individuals misfit the model with high statistics, and 19 misfit with low statistics. For outfit, 11 misfit with high statistics and 16 misfit with low statistics. Analyzing the subscales also revealed instances of person misfit. Twelve percent (n=21) was found to misfit for both infit (high, n=11; low, n=10) and outfit (high, n=11; low, n=10) statistics on the pain subscale. For the disability subscale, 16 individuals (9%) were found to misfit with high (n=12) or low (n=4) infit statistics, and 15 individuals (8%) were found to misfit with high (n=8) or low (n=7) outfit statistics.

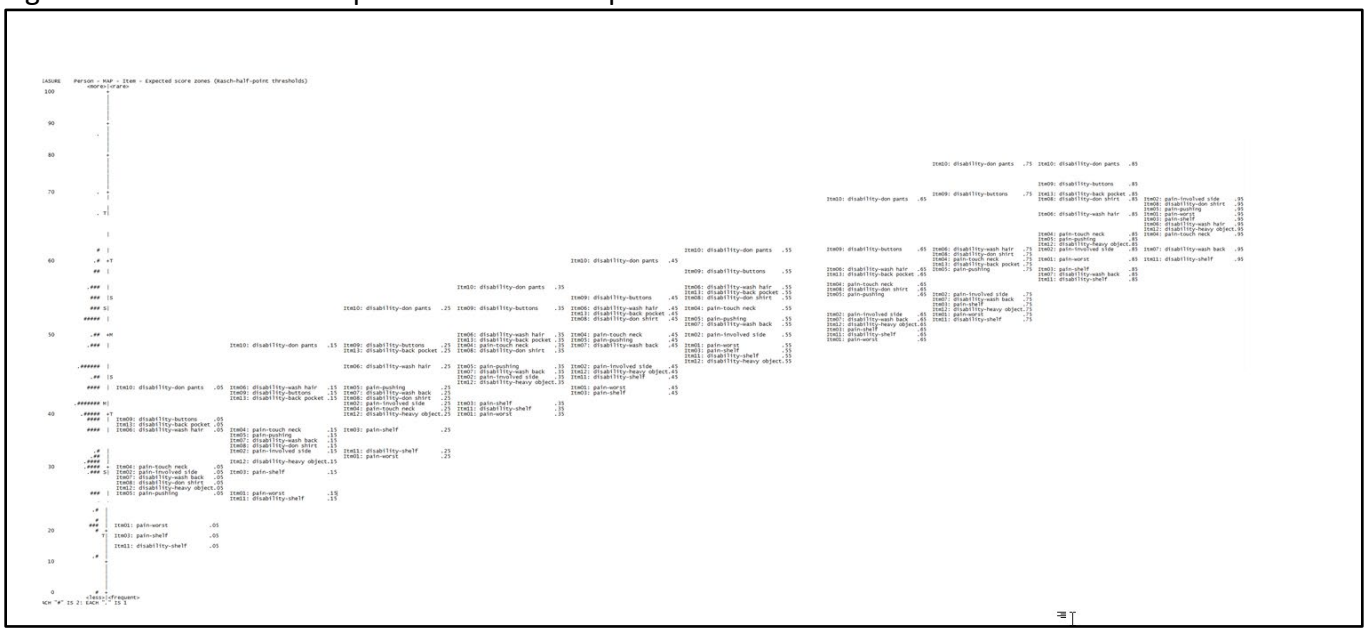
Scale Structure

Scale structure for the SPADI was analyzed using the full scale. Consideration of the Rasch half-point threshold map (Figure 4) demonstrates the scales ability to quantify the majority of ability levels, except for at the extremes. No floor or ceiling effect is present. The map reveals five gaps in the scale structure and 33 occurrences in which more than one response option measured an ability level (redundancy). In this population, Item 10: *disability-don pants* was the easiest item, or the hardest to endorse as being difficult, with a logit (SE) value of 60.8 (0.7). Items 1: *pain-worst* and 11: *disability-shelf* were found to be the most difficult items (easiest to endorse as being difficult) with logit values of 43.9 (0.5) and 43.2 (0.5) respectively.

Response category utilization was analyzed. Each of the 13 test items had one or more response options that did not meet the 10 response requirement (Appendix 9). Response options 9 and 10 were selected least frequently, followed by response options 6, 8 and 7. Uniform distribution of category utilization across the measure was, therefore, not achieved. Issues with disordered step calibration were found in all 13 test items, and failure to achieve monotonic

advancement of observed averages in nine test items. Four test items were found to have high outfit MNSQ values for one or more response categories.

Figure 4. SPADI Rasch Half-point Threshold Map



In an effort to improve category utilization and scale structure, further analysis of the impact of collapsing response options 9 and 10 was performed. Eleven of the 12 test items failed to achieve greater than 10 responses per response option with this modification. Uniform distribution of category utilization remained an issue, and there was no change in the items with high response category outfit statistics. One item improved with respect to monotonicity, however continued to have disordered step calibration. When comparing full scales to collapsed versions, minimal differences were overserved in person (3.66, $\alpha = 0.93$) and item separation (8.26, $\alpha = 0.99$), and item fit and PCA were not impacted. Observation of the Rasch half-point threshold map demonstrated decreased coverage of ability levels at the high (low ability) level of the scale.

Next, response options 8-10 were collapsed. While factors related to scale structure including category utilization (10 test items did not meet criteria) and progression of average

measures improved (six of 11 items did not have expected progression), disordered step calibration increased to seven items, and five items were found to have high response category outfit MNSQ statistics. In addition, item misfit increased to include Item 1: *pain-worst* for outfit, in addition to the previous misfitting Item 12: *disability-heavy object*. Summary statistics for person separation remained stable, however item separation increased to 9.08 (α = 0.99). Observation of the Rasch half-point threshold showed further decrease in coverage of the lower ability levels as well. As a result of this and the increased instability of test items noted with item fit, further analysis of collapsing response options was discontinued.

Reliability

Person separation index (person reliability) for the three SPADI scales are as follows: full 3.74 (α = 0.93); disability 2.71 (α = 0.88); and pain 3.74 (α = 0.93). Each scale meets criteria for adequately classifying individuals into ability levels. Item separation (reliability) indices for all three scales also meet criteria to verify scale hierarchy (full 8.31, α = 0.99; disability 7.56, α = 0.98; and pain 6.51, α = 0.98). Estimates for reliability for person separation in Winsteps is excellent for all three scales (full α = 0.97; disability α = 0.94; and pain α = 0.93).

Neck Dissection Impairment Index

Of note, the NDII response scale is the only reverse-scaled measure included in this study. To allow for accurate interpretation, responses were reversed coded to align the NDII scale with the other PROs being examined. For example, a response option of 5 indicating "not at all" or no difficulty was recoded to represent "a lot" of difficulty. See Appendix 10 for NDII data including: test items, item stems, item logit values, item fit statistics, and response category utilization.

Scale Dimensionality

PCA of the NDII reveals 66.4% of the raw variance is explained by the measure (eigenvalue 2.1). The items that loaded in the positive direction (Items 1: *pain/discomfort* and 2: *stiffness*) relate to symptoms resulting from neck dissection, whereas the items that loaded negatively (Items 4: *lift-light objects* and 5: *lift-heavy objects*) relate to functional ability (Table 11). Given the close proximity of this eigenvalue to the cutoff and the brevity of the scale (10 items) further analysis of potential subscales was not pursued. In addition, no items misfit the model, and DIF was not found to be present for age or gender.

Similar to the other scales, person misfit was also an issue for the NDII. Seventeen percent (n=30) misfit the model with high (n=14) or outfit (n=16) statistics, and 13% (n=24) misfit the model with high (n=10) or low (n=4) outfit statistics.

Table 11. Principal Components Analysis for the NDII

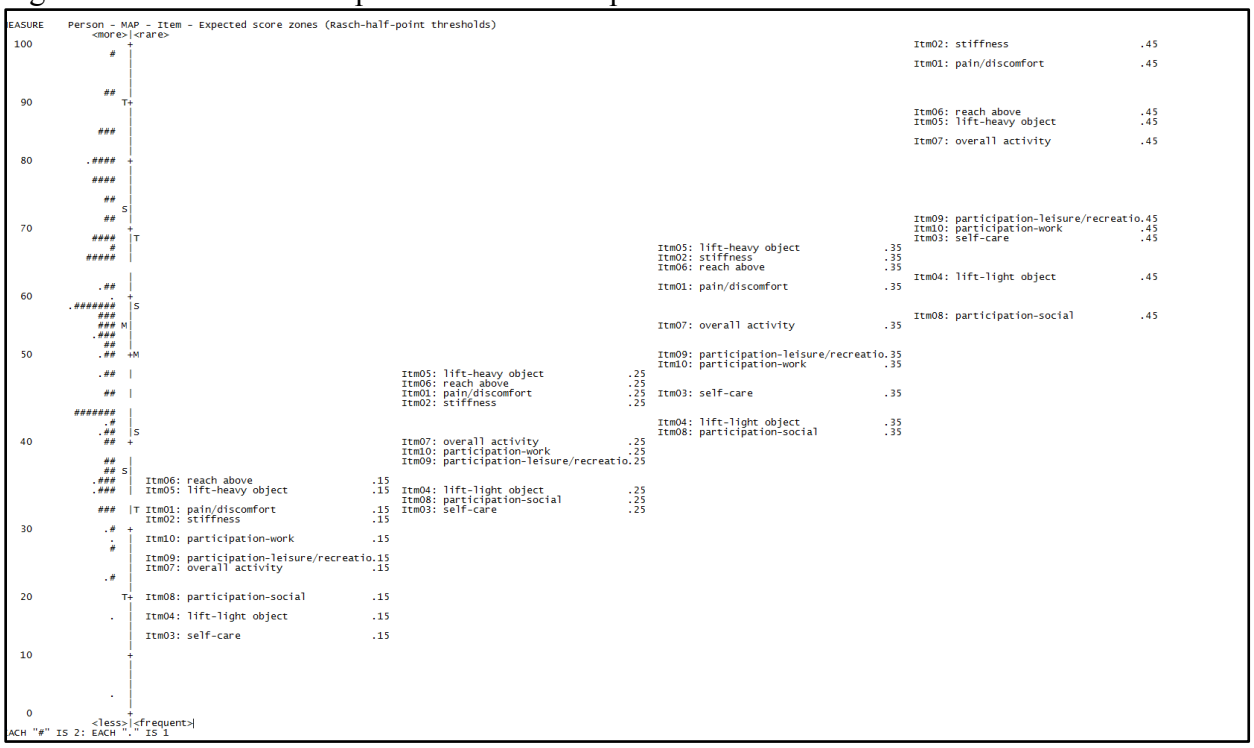
Item	**Loading**
Item 02: *stiffness*	0.84
Item 01: *pain/discomfort*	0.74
Item 04: *lift-light objects*	-0.50
Item 05: *lift-heavy objects*	-0.48

Scale Structure

Observation of the Rasch half-point threshold map (Figure 5) reveals no floor or ceiling effect for this sample. The measure covers most ability levels except for the very low end of the scale (high ability). There were 13 gaps in the measurement scale and two occurrences in which more than one response option targeted the same ability level. Item 8: *participation-social* was found to be the most difficult item (easiest to endorse as being difficult) with a logit value of 38.1 (1.0). Item 2: *stiffness* was found to be the easiest item (most difficult to endorse as being difficult) with a logit value of 62.1 (1.1).

Five of the 10 test items did not meet criteria for 10 or more responses per response option. Items 1- 4 and 8 each had one option with less than 10 responses, three of which were option 1, and two were option 5 (Appendix 10). Because of this inconsistency, further consideration of collapsing the scale structure was not considered. Uniform distribution of response options across categories was not achieved for any test item; conversely the average measure progressed monotonically for each item, there were no issues with disordered step calibration or high outfit MNSQ values for response categories.

Figure 5. NDII Rasch Half-point Threshold Map



Reliability

Summary statistics for the NDII met requirements for population based analysis. Person separation was 3.16 (α = 0.91), and item separation was 8.09 (α = 0.98). Cronbach's alpha for WINSTEPS was excellent (α =0.93).

Summary of the Results

A sample of 182 individuals was recruited for this study. The mean (SD, range) age is 63 years (11.47, 33-87). The majority of the sample is white, non-Hispanic (88%) and male (79%). Participants were on average 174 days (5.8 months) from surgery, however the sample included individuals who were 14 to 567 days (2 weeks to 18 months) from surgery. Most of the cancer diagnoses were of the hypopharynx or oropharynx, and there was equal distribution of individuals receiving surgery only, surgery+radiation, and surgery+radiation+chemotherapy. The majority of cancer care was provided in Arizona and Minnesota, although the sample has good representation of the continental United States. All individuals enrolled, reported some level of shoulder impairment; and only 29% of these individuals had not received some intervention for their symptoms.

Test score means, SD, SEM and 95% CIs are reported for each of the scales. The DASH and QuickDASH are highly correlated as expected, as are the SPADI and its subscales. Significant, although weak, correlations are found between the NDII x DASH, NDII x SPADI, and the UW-QoL (shoulder subscale) x SPADI. Based upon findings of little to no correlation, the UW-QoL (shoulder subscale) was not further considered for usability as a screening tool.

Rasch analysis was utilized to study the validity and reliability of each measure. Scale structure was also considered. A comparison of the four measures is available in Table 12. The DASH did not meet Rasch analysis criteria for unidimensionality, and was deemed inappropriate for utilization in this sample of the population. The QuickDASH, SPADI and NDII were all determined to be unidimensional. The DASH, QuickDASH and SPADI each had issues with item misfit and DIF. All scales had similar issues with person fit, coverage of ability levels per the Rasch half-point threshold map, and gaps and redundancies in scale hierarchy. All scales

failed to meet the 10-response requirement for each response option for each item, and lacked uniform distribution of response category utilization. The DASH, QuickDASH and SPADI also were found to have issues in scale structure related to disordered step calibration and average measures, and high outfit MNSQ values for response categories. The NDII was found to meet most requirements of an optimal rating scale. All measures were found to meet thresholds for person and item separation and reliability statistics.

Table 12. Summary of Findings for the DASH, QuickDASH, SPADI, and NDII

	DASH	**QuickDASH**	**SPADI**	**NDII**
Unidimensional	No	Yes	Yes	Yes
Item misfit	5 items (17%)	2 items (18%)	1 item (8%)	0 items (0%)
Person misfit	40 persons (22%)	16 persons (9%)	36 persons (20%)	31 persons (17%)
DIF (gender)	Not tested	3 items (27%)	1 item (8%)	0 items (0%)
Variable Map	Lacks high disability	Lacks high disability	Lacks coverage at both extremes	Lacks low disability
Logit range	38.7 – 66.5	37.8 – 66.5	43.2 – 60.8	38.1 – 62.1
Floor/ceiling	No	No	No	No
Gaps	9	10	5	13
Redundancies	26	6	33	2
10-Responses?	30%	27%	0%	50%
Uniform distribution?	0%	0%	0%	0%
Disordered step calibration?	27%	18%	100%	0%
Disordered average measures?	17%	0%	69%	0%
Response category outfit MNSQ >2.0?	17%	9%	31%	0%
Person Indices	4.26, $\alpha = 0.95$	2.49, $\alpha = 0.86$	3.74, $\alpha = 0.93$	3.16, $\alpha = 0.91$
Item Indices	7.30, $\alpha = 0.98$	6.60, $\alpha = 0.98$	8.31, $\alpha = 0.99$	8.09, $\alpha = 0.98$
Cronbach's α	$\alpha = 0.96$	$\alpha = 0.90$	$\alpha = 0.97$	$\alpha = 0.93$

Chapter 5: Discussion

Introduction to the Chapter

This section provides an in-depth discussion of the study results and outcomes of the study aims outlined in Chapter 1. Clinical implications, recommendations, and limitations of this study are also discussed.

Discussion

This study sought to answer the question, "Which of the recommended patient-reported outcome measures (PRO) demonstrates acceptable psychometric characteristics allowing for accurate test score interpretation in patients presenting with shoulder disability in the setting of head and neck cancer (HNC)?" The aims of the study were three-fold: (1) to assess, using Rasch analysis, the construct validity and overall appropriateness of test score interpretation of the Disability of the Arm, Shoulder and Hand (DASH), QuickDASH, Shoulder Pain and Disability Index (SPADI) and the Neck Dissection Impairment Index (NDII); (2) determine the appropriateness of use of the University of Washington Quality of Life (UW-QoL) shoulder subscale as a screening tool for shoulder-related impairment; and (3) suggest a combination of outcome measures or a new measure that most accurately portrays shoulder disability in this population. Through completion of these aims, the 2014 recommendations[2] for outcome measures to quantify shoulder impairment in the HNC population would either be validated or challenged. In a 2014 systematic review, Eden and colleagues strongly recommended the following measures: DASH, SPADI, NDII and the UW-QoL (shoulder subscale), and recommended the QuickDASH.[2] The results of this study strongly support the use of the NDII in this population. The SPADI, and the QuickDASH are recommended, and the DASH is not recommended. It is also not recommended that the shoulder subscale of the UW-QoL be utilized

as a single-item measure to quantify shoulder dysfunction, or to predict performance on the other shoulder-related PROs.

Analysis of the four PROs through a combination of Rasch methodologies and qualitative analysis, shows that the NDII meets assumptions for unidimensionality, covers the majority of ability levels, does not have floor and ceiling effects, has appropriate person and item summary statistics, and meets most requirements for an optimal rating scale as described by Linacre.[182] Weaknesses of the NDII are primarily within the scale hierarchy and scale structure. Specifically, the measure lacks test items to adequately capture individuals with lower disability levels, and has 13 gaps in the rating scale structure, suggesting low scale sensitivity. This is not surprising given that this is the shortest of the four PROs included in this analysis. The results of this study vary from previous reports of Rasch analysis of the NDII, which found disordered response option step categories and floor effects.[21]

Like the DASH, QuickDASH and SPADI, this population failed to achieve the required 10 responses per response option for optimal analysis in the NDII, and therefore lacked uniform utilization of response options. Because this was a consistent finding in this study, it is assumed that this sample of the HNC population did not experience the level of disability that would warrant utilization of the response options indicating high disability.

The QuickDASH and the SPADI were found to have some flaws which suggest that the user proceed with caution when using these scales in the HNC population. Both scales were deemed unidimensional, with minor issues with item misfit and DIF. Floor and ceiling effects do not exist for either scale. The QuickDASH does not possess test items to adequately capture individuals with high disability, whereas the SPADI fails to capture individuals at the extremes of ability levels. Both have gaps and item redundancies. The SPADI has fewer gaps, and more

redundancies due to the greater number of test items and response options inherent in the measure. Both scales also fail to meet all five criteria for optimal rating scales as outlined by Linacre.[182] The SPADI has superior item and person summary statistics than the QuickDASH.

QuickDASH scale structure varies slightly than in a sample of individuals who underwent surgery for musculoskeletal causes including total shoulder arthroplasty or rotator cuff repair.[224] Individuals with shoulder impairment due to HNC are least challenged by Item 5: *cut food*, whereas the previously mentioned population is least impacted by Item 10: *tingling*. It is interesting that this was not the case for the HNC population, given that these individuals rarely experience this impairment, and that Item 10: *tingling* was found to misfit the model in this study. Future modifications to the QuickDASH could consider elimination of Item 10: *tingling*, because it is not a common complaint in the HNC population. Both populations were most challenged by Item 6: *recreation–force/impact*. Item 1: *open jar* also misfit the model. Qualitative consideration as to why misfit would occur in this sample is less clear. This task occurs with the arm by the side, which is not a common limitation for individuals with SAN palsy. Removal of this item for further analysis did not impact results. It is possible that this Item 1: *open jar* misfit based upon the presence of DIF by gender, specifically females finding this item to be more difficult than males. Females also found Item 3: *carry bag* to be more difficult than males, whereas males found Item 4: *wash back* to be more difficult than females.

Results support the SPADI as a unidimensional scale as outlined in the literature,[305,314] suggesting that for test score interpretation utilization of the total score, rather than the pain and disability subscores, may be sufficient. Findings of no floor or ceiling effect contradict a previously reported study which found a floor effect of 17% one to three months following neck dissection surgery and a 43% six to eight months following surgery.[302] This study also found

some variation in the item hierarchy. Prior studies report Items 7: *disability-wash back*, 11: *disability-shelf*, and 12: *disability-heavy object* as being the most difficult.[109] In this sample, Items 1: *pain-worst* and 11: *disability-shelf* were found to be the most difficult items, and Item 12: *disability-heavy object* was found to misfit the model. Interestingly, DASH Item 11: *carry object* did not misfit the DASH model. It is possible that this test item is answered differently based upon time since surgery. Individuals who are within the first two to three weeks of recovery from surgery, have a lifting restriction of 10 pounds due to risk of post-operative bleeding.

While the DASH has been widely used in the literature, and recommended for use in the HNC population,[3,66,190] Rasch analysis in this study did not support its use. This is consistent with recent findings in pilot data (Eden, Kunze, Cheng, unpublished data, 2018). Primary limitations of this measure include failure to satisfy the requirement of unidimensionality, and as found in this study, no clear evidence supporting delineated subscales for analysis. Exploratory factor analysis was completed to further consider the presence of additional constructs for consideration as recommended in the literature,[14] without clear constructs identified. Similar to prior reports, the DASH was found to have item misfit, and gaps and redundancies in the scale hierarchy. Additional inconsistencies with the literature were uncovered with respect to item difficulty. Rogers and colleagues found Items 6: *object overhead*, 13: *style hair* and 15: *don sweater* to be the most difficult items; whereas Items 18: *recreation-force/impact* and 27: *weakness* were most difficult in this sample. In a general musculoskeletal population, Item 18: *recreation-force/impact* has been identified, along with Item 21: *sexual activities* and 20: *manage transportation*, as items needing further consideration in future iterations of the DASH.[247] Based upon the findings that Item 18: *recreation-force/impact* is the most difficult

item for the HNC population, it is recommend that this item be retained based upon its value in capturing individuals with higher ability levels. This study supports removal of Item 21: *sexual activities* based upon issues with item misfit. Similar recommendations have been made by other authors [111,205] Of note, Kennedy and colleagues reported problems with Item 1: *open jar*, which was also problematic in this study for both the DASH and QuickDASH.[247] This study uncovered flaws in the scale structure as defined by Linacre.[182] Because of these findings, it was determined that the DASH score cannot be adequately interpreted for this population of patients and a more in-depth analysis of the measure, including consideration of DIF, alternate scale structures, or subscales, was not pursued. This research study supports the use of the QuickDASH over the DASH. Similar to the reports of MacDermid and colleagues, the QuickDASH covers the same range of ability levels as the DASH.[224]

The recommendation not to use the UW-QoL (shoulder subscale) as a screening tool in the HNC population varies from previous recommendations.[290,345] One reason as to why our recommendations may vary from Rogers is that our sample had very low utilization of response options 1 and 4. Response option 1 ("I have no problems with my shoulder") was rarely utilized (n = 3) because the inclusion criteria for the study required subjects to report some degree of shoulder impairment to be eligible for participation. Of significance is the little to no correlation with the other PROs in this study. The UW-QoL (shoulder subscale) was not found to have a statistically significant correlation with the DASH, QuickDASH or the NDII, however a weak, negative correlation was found with the single-item question and the SPADI. Of interest, the DASH, QuickDASH, SPADI and UW-QoL (shoulder subscale) have the same scale structure, so any correlation should be in the positive direction if present. In this case, the UW-QoL item did

not behave as expected. Future studies should consider the reliability and construct validity of the UW-QoL scale in its entirety using Rasch analysis.

The lack of statistically significant correlation between the DASH/QuickDASH and the SPADI, and the weak correlation with the NDII and the DASH, and the NDII and the SPADI, were unexpected findings in this analysis. Our results found significant deviations from published findings for relationships between the other PROs included in this study. For instance, the DASH and NDII have been reported to have a strong negative relationship in the HNC population ($r = -0.86$),[196] whereas our results found a weak, although significant, correlation. The NDII has also previously been found to have a strong correlation with the shoulder subscale of the UW-QoL ($r = 0.75$),[290] and SPADI ($r = -0.75$)[21] which was not the case in this study.

All PROs lacked the requirements for an optimal rating scale as defined by Linacre.[182] Most notable in this sample was the failure to achieve a minimum of 10 responses per response option on each test item. Particular attention was paid to this through study recruitment. Although this sample includes a wide variety of HNC-related tumors, surgeries, adjuvant treatments, and time from surgery, we were unable to enroll subjects with very high levels of shoulder disability, as noted by low utilization of the higher response options across measures. A consideration for future studies would be to focus on individuals with higher levels of disability, although the incidence of this occurring is decreasing as surgical techniques improve. Another related limitation of scale structure across the measures is related to uniform distribution of responses across response options, which again is likely related to the sample and the captured ability levels.

To address the limitation of low utilization of response options, this study included trials of collapsing response categories for the QuickDASH and the SPADI. Collapsing response

options was not performed for the NDII because a clear pattern of poor utilization did not exist. Future use of the PROs in the HNC population could consider collapsing response options 4 and 5 in the QuickDASH. Upon collapsing these response options, response option utilization improved, as did the presence of disordered step calibration and average measures, and high response category outfit statistics. Additionally, scale summary statistics and PCA findings did not change significantly. Options for the SPADI require further analysis, however. Collapsing options 9 and 10 did not adequately address limitations in scale structure, and with a trial of combining response options 8, 9 and 10, the scale structure further deteriorated.

Another similar theme across measures is the presence of person misfit, which may have contributed to item misfit in the DASH, QuickDASH, and the SPADI. Person misfit was consistent across measures and therefore was attributed to the sample for this study. Given the absence of item misfit in the NDII, and the already low response category utilization for all measures a qualitative analysis of the misfitting persons for consideration of removal of misfitting persons was not performed.

The findings in this study may be related to variations between the PROs, including purpose, test item construction and recall period. The DASH and QuickDASH measure disability related to impairment of the entire upper extremity, whereas the SPADI measures disability and pain specific to the shoulder. Except for situations of radial forearm free flap harvest for surgical reconstruction, the entire upper extremity is rarely affected in the medical management of HNC. The NDII, a measure of health-related quality of life (HRQOL) due to shoulder impairment is the only measure included in the Rasch analysis designed specifically for the HNC population. It is also the only measure included in this analysis that includes questions related to the neck. Individual test items within each measure are written differently to capture the intent of the PRO.

Test items in the DASH/QuickDASH and SPADI are worded similarly and request the responder to indicate how much difficulty an activity would cause. The NDII test items address the same activities however each test item relates the impairment back to the shoulder or neck. In addition the NDII test items appear more personal because they address the respondent through the use of "you." For example, Item 1 reads "Are you bothered by neck or shoulder pain or discomfort?" It is possible that the differences in test item construction, and specificity of the NDII to the HNC population impacted how the study participants answered, and therefore how the data fit the Rasch model. Another inherent difference in the scales is the recall period. The DASH/QuickDASH, SPADI and the UW-QoL (shoulder subscale) ask respondents to consider the past week when answering test items, whereas the NDII recall period is four weeks.

Given the satisfactory performance of the SPADI, NDII and QuickDASH in this population, consideration of a combination of PROs for a new measure is not necessary. The idea not to create a new PRO is supported by the notion that there will never be one "perfect" measure. The availability of too many PROs can lead to confusion and limit the ability to compare outcomes in clinical or research settings.[351] It is also not recommended that a therapist utilize more than one of the recommend PROs on a single patient. Consideration of individual test items within each scale reveals redundancy in content, and therefore would likely introduce redundancy in test items available to quantify ability levels. An attempt to decrease gaps in item difficulty through the use of two more scales would introduce unnecessary responder burden. Therefore the third aim of this study was not pursued.

Implications

This study has significant clinical implications for medical providers working with individuals with HNC. Individuals receiving a neck dissection procedure for their diagnosis are

at increased risk for experiencing shoulder impairment, and decreased QOL. Referral to a physical therapist can help to minimize impairment. Utilization of PROs can provide a baseline level of function prior to surgery, or show progress of return to function following surgery. As payment models change to performance-based models, healthcare providers must be able to show value in the services provided through the utilization of valid and reliable instruments. This study provides valuable insight into previously recommended PROs for this population.[3,66,190] The most commonly utilized and recommended PROs, DASH, QuickDASH, and SPADI, have flaws which limit test score interpretation in this population. In addition, we have shown that the single-item shoulder question is not appropriate for use as a screening tool for this sample of the population. A strength of this study is the validation of the NDII as a measure of shoulder-related impairment in the HNC population.

Recommendations

The results of this study strongly support the use of the NDII. The SPADI, and the QuickDASH are recommended with reservation, and the DASH is not recommended. It is also not recommended that the shoulder subscale of the UW-QoL be utilized as a single-item measure to quantify shoulder dysfunction, or to predict performance on the other shoulder-related PROs. The EDGE Task Force recommendations[2] have not been fully supported, therefore an updated recommendation should be made based upon the findings of this study. Future studies should further strive to validate the findings in this study related to correlational analysis between measures. Research is also needed to provide recommendations on alternative scale structures, related to collapsing response options, for the QuickDASH and the SPADI. Additional research should consider the responsiveness of the NDII in clinical settings.

Limitations and Delimitations

Limitations

The primary limitation in this study is with respect to the sample of subjects enrolled, which contributed to issues of person misfit, underutilized response categories and lack of uninform distribution of category utilization for each PRO, and in some scales possibility of multiple constructs. The sample had high incidence of person misfit, indicating that 14-20% of the sample answered test questions in a way that was not predicted by the model. It is possible that the higher incidence of person misfit is because of our attempt to utilize PROs developed for a musculoskeletal population and apply this to a HNC population, however the lowest incidence of person misfit was found in the QuickDASH, and the highest in the DASH. If our hypothesis was correct, the NDII would have the lowest incidence. Other possibilities for high person misfit could be related to responder burden and guessing. Each participant was asked to complete a packet of documents which required quite a bit of reading and completion of a total of four questionnaires (greater than 60 test questions). The packets were provided to each participant in the same order, therefore if responder burden is present, we would expect to see higher person misfit in the PROs later in the packet (SPADI and NDII), which again was not the case.

An additional limitation of this research study was the failure to obtain a heterogeneous sample of sufficient ability range to accurately assess the response scale structure and item calibration for each of the PROs. This occurred despite attempts to include subjects with greater disability levels including those undergoing bilateral neck dissections and those who's SAN(s) was sacrificed during surgery. Pilot data reflected a decreased tendency for utilization of response options (4 and 5) which suggest greater disability if endorsed on the DASH and QuickDASH.[19] This was confirmed again in this study, and was also found in the SPADI and the

some items in the NDII. In addition, in most cases the PROs failed to achieve uniform distribution of category utilization. This is also likely sample-related. It is possible that the HNC population does not reach the necessary levels of disability to utilize the response options indicating lower ability levels.

Delimitations

Delimitations are factors that impact the research study which are under the PI's control. Delimitations for this study included slower than anticipated subject enrollment, incomplete or missing data, and minor errors in data entry. Slower than anticipated subject enrollment was the primary limitation and resulted from decreased awareness, decreased motivation, and time constraints of a busy clinical calendar for medical providers asked to participate in the study. This was addressed successfully through utilization of mailed questionnaires. Incomplete or missing data was present, but limited by regular communication with study staff for quality control during data entry. Because of the nature of mailed questionnaires it was impossible to ensure completeness of PROs, however care was taken to review accuracy in those that were finished in a one-on-one setting. Additionally, data coordinators and the primary investigator completed data checks to verify accuracy of data in REDcap database.

Summary

This chapter provides a discussion of the study results , limitations, and recommendations. Rasch analysis results indicate that the NDII is the most appropriate measure studied for use in a population of patients with HNC reporting shoulder impairment resulting from neck dissection procedure. The QuickDASH and SPADI are appropriate but do have some limitations. The DASH is not recommended, neither is the UW-QoL (shoulder subscale) as a

screening tool. Future research could address alternate scale structures, including collapsing response options, and further analysis of the correlational relationships between the PROs.

Reference List

1. Ewing MR, Martin H. Disability following "radical neck dissection:" An assesment based on the postoperative evaluation of 100 patients. *Cancer.* 1952;5:873-883.
2. Eden MM, Flores AM, Galantino ML, Spinelli BA. Recommendations for patient-reported outcome measures for head and neck cancer-related shoulder dysfunction: a systematic review. *Rehabil Oncol.* 2014;32(3):6-19.
3. Goldstein DP, Ringash J, Bissada E, et al. Evaluation of shoulder disability questionnaires used for the assessment of shoulder disability after neck dissection for head and neck cancer. *Head Neck.* 2014;36:1453-1458.
4. Taylor RJ, Chepeha JC, Teknos TN, et al. Development and validation of the neck dissection impairment index: a quality of life measure. *Arch Otolaryngol Head Neck Surg.* 2002;128(1):44-49.
5. Pusic A, Liu JC, Chen CM, et al. A systematic review of patient-reported outcome measures in head and neck cancer surgery. *Otolaryngol Head Neck Surg.* 2007;136(4):525-535.
6. Dye DC, Eakman AM, Bolton KM. Assessing the validity of the Dynamic Gait Index in a balance disoders clinic: An application of Rasch Analysis. *Phys Ther.* 2013;93(6):809-818.
7. Chiu Y-P, Fritz SL, Light KE, Velozo CA. Use of Item Response Analysis to Investigate Measurement Properties and Clinical Validity of Data for the Dynamic Gait Index. *Phys Ther.* 2006;86:778-787.
8. Wong CK, Chen CC, Welsh J. Preliminary assessment of balance with the berg balance scale in adults who have a leg amputation and dwell in the community: Rasch rating scale analysis. *Phys Ther.* 2013;93(11):1520-1529.
9. Straube D, Moore J, Leech K, Hornby TG. Item analysis of the berg balance scale in individuals with subacute and chronic stroke. *Top Stroke Rehabil.* 2013;20(3):241-249.
10. Franchignoni F, Ferriero G, Giordano A, Sartorio F, Vercelli S, Brigatti E. Psychometric properties of QuickDASH - a classical test theory and Rasch analysis study. *Man Ther.* 2011;16(2):177-182.
11. Cano S, Barrett L, Zajicek J, Hobart J. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. *Mult Scler.* 2011;17(2):214-222.
12. Di Pietro F, Catley MJ, McAuley JH, et al. Rasch analysis supports the use of the Pain Self-Efficacy Questionnaire. *Phys Ther.* 2014;94(1):91-100.
13. Forget NJ, Jerosch-Herold C, Shepstone L, Higgins J. Psychometric evaluation of the Disabilities of the Arm, Shoulder and Hand (DASH) with Dupuytren's contracture: validity evidence using Rasch modeling. *BMC Musculoskelet Disord.* 2014;15(361).
14. Franchignoni F, Giordano A, Sartorio F, Vercelli S, Passcariello B, Ferriero G. Suggestions for refinement of the Disabilities of the Arm, Shoulder and Hand Outcome Measure (DASH): a factor analysis and Rasch validation study. *Arch Phys Med Rehabil.* 2010;91(9):1370-1377.
15. Lehman LA, Woodbury M, Velozo CA. Examination of the Factor Structure of the Disabilities of the Arm, Shoulder, and Hand Questionnaire. *Am J Occup Ther.* 2011;65(2):169-178.

16. Tyser AR, Beckmann J, Franklin JD, et al. Evaluation of the PROMIS physical function computer adaptive test in the upper extremity. *J Hand Surg Am.* 2014;39(10):2047-2051.e2044.
17. Braitmayer K, Dereskewitz C, Oberhauser C, Rudolf KD, Coenen M. Examination of the Applicability of the Disabilities of the Arm, Shoulder and Hand (DASH) Questionnaire to Patients with Hand Injuries and Diseases Using Rasch Analysis. *Patient.* 2017;10(3):367-376.
18. Dalton E, Lannin NA, Laver K, et al. Validity, reliability and ease of use of the disabilities of arm, shoulder and hand questionnaire in adults following stroke. *Disabil Rehabil.* 2016:1-8.
19. Eden M, Straube D. Rasch Analysis of the Disabilities of the Arm, Shoulder and Hand and the QuickDASH in Patients Undergoing Neck Dissection for Treatment of Head and Neck Cancer (poster presentation). Paper presented at: American Physical Therapy Association Combined Sections Meeting2015; Indianapolis, Indiana.
20. Cook KF, Gartsman GM, Roddey TS, Olson SL. The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation. *Arch Phys Med Rehabil.* 2001;82(11):1558-1565.
21. Stuiver MM, ten Tusscher MR, van Opzeeland A, et al. Psychometric properties of three patient reported outcome measures for the assessment of shoulder disability after neck dissection. *Head Neck.* 2016;38(1):102-110.
22. American Physical Therapy Association. Functional Limitation Reporting Under Medicare. http://www.apta.org/payment/medicare/codingbilling/functionallimitation/. Accessed August 3, 2013.
23. National Cancer Institute. Head and Neck Cancers. http://www.cancer.gov/cancertopics/factsheet/Sites-Types/head-and-neck. Accessed January 1, 2015.
24. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J Clin.* 2014;64(1):9-29.
25. Chaturvedi AK, Engels EA, Pfeiffer RM, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J Clin Oncol.* 2011;29(32):4294-4301.
26. Deschler DG, Richmon JD, Khariwala SS, Ferris RL, Wang MB. The "New" Head and Neck Cancer Patient-Young, Nonsmoker, Nondrinker, and HPV Positive: Evaluation. *Otolaryngol Head Neck Surg.* 2014.
27. D'Souza G, Dempsey A. The role of HPV in head and neck cancer and review of the HPV vaccine. *Prev Med.* 2011;53:S5-S11.
28. Courneya KS. Exercise in cancer survivors: an overview of research. *Med Sci Sports Exerc.* 2003;35(11):1846-1852.
29. American Cancer Society. Cancer Facts and Figures 2014. 2014; http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf. Accessed April 19, 2014.
30. Holmes JD. Neck dissection: nomenclature, classification, and technique. *Oral Maxillofac Surg Clin North Am.* 2008;20(3):459-475.
31. Robbins KT, Shaha AR, Medina JE, et al. Consensus Statement on the Classification and Terminology of Neck Dissection. *Arch Otolaryngol Head Neck Surg.* 2008;134(5):536-538.

32. Crile G. Excision of cancer of the head and neck. *JAMA.* 1906;47:1780-1786.
33. Kierner AC, Zelenka I, Heller S, Burian M. Surgical anatomy of the spinal accessory nerve and the trapezius branches of the cervical plexus. *Arch Surg.* 2000;135:1428-1431.
34. Brown H. Anatomy of the Spinal Accessory Nerve Plexus: Relevance to Head and Neck Cancer and Atherosclerosis. *Exp Biol Med.* 2002;227:570-578.
35. Kierner AC, Burian M, Bentzien S, Gstoettner W. Intraoperative electromyography for identification of the trapezius muscle innervation: clinical proof of a new anatomical concept. *Laryngoscope.* 2002;112(10):1853-1856.
36. Saunders JR, Jr., Hirata RM, Jaques DA. Considering the spinal accessory nerve in head and neck surgery. *Am J Surg.* 1985;150(4):491-494.
37. Umeda M, Shigeta T, Takahashi H, et al. Shoulder mobility after spinal accessory nerve-sparing modified radical neck dissection in oral cancer patients. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod.* 2010;109(6):820-824.
38. Güldiken Y, Orhan KS, Demirel T, Ural Hİ, Yücel EA, Değer K. Assessment of shoulder impairment after functional neck dissection: Long term results. *Auris Nasus Larynx.* 2005;32(4):387-391.
39. Speksnijder CM, van der Bilt A, Slappendel M, de Wijer A, Merkx MAW, Koole R. Neck and shoulder function in patients treated for oral malignancies: A 1-year prospective cohort study. *Head Neck.* 2012:n/a-n/a.
40. Remmler D, Byers R, Scheetz J, et al. A Prospective Study of Shoulder Disability Resulting from Radical and Modified Neck Dissections. *Head Neck Surg.* Mar/Apr 1986;8:280-286.
41. Leipzig B, Suen JY, English JL, Barnes J, Hooper M. Functional evaluation of the spinal accessory nerve after neck dissection. *Am J Surg.* 1983;146(4):526-530.
42. Stacey RJ, O'Leary ST, Hamlyn PJ. The innervation of the trapezius muscle: a cervical motor supply. *J Craniomaxillofac Surg.* 1995;23(4):250-251.
43. Walker H. Cranial Nerve XI: The Spinal Accessory Nerve. In: Walker H, Hall W, Hurst J, eds. *Clinical Methods: The History, Physical, and Laboratory Examinations.* 3rd ed. Boston: Butterworths; 1990.
44. Lee SH, Lee JK, Jin SM, et al. Anatomical variations of the spinal accessory nerve and its relevance to level IIb lymph nodes. *Otolaryngol Head Neck Surg.* 2009;141(5):639-644.
45. Restrepo CE, Tubbs RS, Spinner RJ. Expanding what is known of the anatomy of the spinal accessory nerve. *Clin Anat.* 2014.
46. Canella C, Demondion X, Abreu E, Marchiori E, Cotten H, Cotten A. Anatomical study of spinal accessory nerve using ultrasonography. *Eur J Radiol.* 2013;82(1):56-61.
47. Bater MC, Dufty J, Brennan PA. High division of the accessory nerve: a rare anatomical variation as a possible pitfall during neck dissection surgery. *J Craniomaxillofac Surg.* 2005;33(5):340-341.
48. Saman M, Etebari P, Pakdaman MN, Urken ML. Anatomic relationship between the spinal accessory nerve and the jugular vein: a cadaveric study. *Surg Radiol Anat.* 2011;33(2):175-179.
49. Alemanno F, Egarter Vigl E. Anatomy of the Cervical Plexus. In: *Anesthesia of the Upper Limb: A State of the Art Guide.* Milan: Springer; 2014:37-40.
50. Pu YM, Tang EY, Yang XD. Trapezius muscle innervation from the spinal accessory nerve and branches of the cervical plexus. *Int J Oral Maxillofac Surg.* 2008;37(6):567-572.

51. Tubbs RS, Shoja MM, Loukas M, et al. Study of the cervical plexus innervation of the trapezius muscle. *J Neurosurg Spine.* 2011;14(5):626-629.
52. Krause HR. Reinnervation of the trapezius muscle after radical neck dissection. *J Craniomaxillofac Surg.* 1994;22(6):323-329.
53. Garzaro M, Riva R, Raimondo L, Aghemo L, Ciordano C, Pecorari G. A study of neck and shoulder morbidity following neck dissection: The benefits of cervical plexus preservation. *Ear Nose Throat J.* 2015;94(8):330-344.
54. Agur AMR, Lee MJ. Upper Limb. In: *Grant's Atlas of Anatomy.* 10th ed. Philadelphia: Lippincott Williams & Wilkins; 1999.
55. Wiater J, Bigliani L. Spinal Accessory Nerve Injury. *Clin Orthop Relat Res.* 1999;368:5-16.
56. Herring D, King AI, Connelly M. New rehabilitation concepts in management of radical neck dissection syndrome. A clinical report. *Phys Ther.* 1987;67(7):1095-1099.
57. Cappiello J, Piazza C, Nicolai P. The spinal accessory nerve in head and neck surgery. *Curr Opin Otolaryngol Head Neck Surg.* 2007;15(2):107-111 110.1097/MOO.1090b1013e3280523ac3280525.
58. Hovorka MS, Uray NJ. Microscopic clusters of sensory neurons in C1 spinal nerve roots and in the C1 level of the spinal accessory nerve in adult humans. *Anat Rec (Hoboken).* 2013;296(10):1588-1593.
59. Brown H, Burns S, Kaiser CW. The spinal accessory nerve plexus, the trapezius muscle, and shoulder stabilization after radical neck cancer surgery. *Ann Surg.* 1988;208(5):654-661.
60. Bigliani LU, Compito CA, Duralde XA, Wolfe IN. Transfer of the levator scapulae, rhomboid major, and rhomboid minor for paralysis of the trapezius. *J Bone Joint Surg Am.* 1996;78(10):1534-1540.
61. Dilber M, Kasapoglu F, Erisen L, Basut O, Tezel I. The relationship between shoulder pain and damage to the cervical plexus following neck dissection. *Eur Arch Otorhinolaryngol.* 2007;264(11):1333-1338.
62. van Wilgen CP, Dijkstra PU, van der Laan BFAM, Plukker JT, Roodenburg JLN. Shoulder complaints after neck dissection; is the spinal accessory nerve involved? *Br J Oral Maxillofac Surg.* 2003;41(1):7-11.
63. Tsuji T, Tanuma A, Onitsuka T, et al. Electromyographic findings after different selective neck dissections. *Laryngoscope.* 2007;117(2):319-322.
64. Nori S, Soo KC, Green RF, Strong EW, Miodownik S. Utilization of intraoperative electroneurography to understand the innervation of the trapezius muscle. *Muscle Nerve.* 1997;20(3):279-285.
65. Watkins JP, Williams GB, Mascioli AA, Wan JY, Samant S. Shoulder function in patients undergoing selective neck dissection with or without radiation and chemotherapy. *Head Neck.* 2011;33(5):615-619.
66. Carr SD, Bowyer D, Cox G. Upper limb dysfunction following selective neck dissection: A retrospective questionnaire study. *Head Neck.* 2009;31(6):789-792.
67. St. Louis Children's Hospital, Washington University School of Medicine. Classification of Nerve Injuries. http://brachialplexus.wustl.edu/injury.html. Accessed January 4, 2015.
68. Mackinnon SE, Novak CB. Nerve Injury & Recovery. 2001; http://nerve.wustl.edu/NerveInjury.pdf. Accessed January 4, 2015.

69. Fawcett JW, Keynes RJ. Peripheral Nerve Regeneration. *Annu Rev Neurosci.* 1990;13:43-60.
70. Bradley PJ, Ferlito A, Silver CE, et al. Neck treatment and shoulder morbidity: Still a challenge. *Head Neck.* 2011;33(7):1060-1067.
71. Orhan KS, Demirel T, Baslo B, et al. Spinal accessory nerve function after neck dissections. *J Laryngol Otol.* 2007;121(1):44-48.
72. Dijkstra PU, van Wilgen PC, Buijs RP, et al. Incidence of shoulder pain after neck dissection: A clinical explorative study for risk factors. *Head Neck.* 2001;23(11):947-953.
73. Laverick S, Lowe D, Brown JS, Vaughan ED, Rogers SN. The impact of neck dissection on health-related quality of life. *Arch Otolaryngol Head Neck Surg.* 2004;130(2):149-154.
74. Merve A, Mitra I, Swindell R, Homer JJ. Shoulder morbidity after pectoralis major flap reconstruction for head and neck cancer. *Head Neck.* 2009;31(11):1470-1476.
75. Eickmeyer SM, Walczak CK, Myers KB, Lindstrom DR, Layde P, Campbell BH. Quality of Life, Shoulder Range of Motion, and Spinal Accessory Nerve Status in 5-Year Survivors of Head and Neck Cancer. *PMR.* 2014.
76. Steven O. Short M, Jory N. Kaplan M, George E. Laramore M, Carles W.Cummings M. Shoulder Pain and Function After Neck Dissection With or Without Preservation of the Spinal Accessory Nerve. *Am J Surg.* October 1984;148:478-482.
77. Chaplin JM, Morton RP. A prospective, longitudinal study of pain in head and neck cancer patients. *Head Neck.* 1999;21(6):531-537.
78. Teymoortash A, Hoch S, Eivazi B, Werner JA. Postoperative morbidity after different types of selective neck dissection. *Laryngoscope.* 2010;120(5):924-929.
79. Selcuk A, Selcuk B, Bahar S, Dere H. Shoulder function in various types of neck dissection: role of spinal accessory nerve and cervical plexus preservation. *Tumori.* Jan-Feb 2008;94(1):36-39.
80. Erisen L, Basel B, Irdesel J, et al. Shoulder function after accessory nerve-sparing neck dissections. *Head Neck.* 2004;26(11):967-971.
81. Schuller DE, Reiches NA, Hamaker RC, et al. Analysis of disability resulting from treatment including radical neck dissection or modified neck dissection. *Head Neck Surg.* 1983;6:551-558.
82. van Wouwe M, de Bree R, Kuik DJ, et al. Shoulder morbidity after non-surgical treatment of the neck. *Radiother Oncol.* 2009;90(2):196-201.
83. Kuntz AL, Weymuller EA, Jr. Impact of neck dissection on quality of life. *Laryngoscope.* 1999;109(8):1334-1338.
84. Stubblefield MD. Radiation Fibrosis Syndrome: Neuromuscular and Musculoskeletal Complications in Cancer Survivors. *PMR.* 2011;3(11):1041-1054.
85. Nowak P, Parzuchowski J, Jacobs JR. Effects of combined modality therapy of head and neck carcinoma on shoulder and head mobility. *J Surg Oncol.* 1989;41(3):143-147.
86. Chan JY, Wong ST, Chan RC, Wei WI. Shoulder Dysfunction after Selective Neck Dissection in Recurrent Nasopharyngeal Carcinoma. *Otolaryngol Head Neck Surg.* 2015;153(3):379-384.
87. Gane EM, O'Leary SP, Hatton AL, Panizza BJ, McPhail SM. Neck and Upper Limb Dysfunction in Patients following Neck Dissection: Looking beyond the Shoulder. *Otolaryngol Head Neck Surg.* 2017:194599817721164.

88. Zackrisson B, Mercke C, Strander H, Wennerberg J, Cavallin-Stahl E. A systematic overview of radiation therapy effects in head and neck cancer. *Acta Oncol.* 2003;42(5-6):443-461.
89. World Health Organization. International Classification of Functioning, Disability and Health (ICF). http://www.who.int/classifications/icf/en/. Accessed May 31, 2014.
90. Goldstein DP, Ringash J, Bissada E, et al. Scoping review of the literature on shoulder impairments and disability after neck dissection. *Head Neck.* 2013.
91. Nahum AM, Mullally W, Marmor L. A syndrome resulting from radical neck dissection. *Arch Otolaryngol.* 1961;74:424-428.
92. Witt RL, Gillis T, Pratt R, Jr. Spinal accessory nerve monitoring with clinical outcome measures. *Ear Nose Throat J.* 2006;85(8):540-544.
93. Cheng PT, Lin YH, Hao SP, Yeh ARM. Objective Comparison of Shoulder Dysfunction After Three Neck Dissection Techniques. *Am Otol Rhinol Laryngol.* 2000;109:761-766.
94. Sheikh A, Shallwani H, Ghaffar S. Postoperative shoulder function after different types of neck dissection in head and neck cancer. *Ear Nose Throat J.* 2014;93(4-5):E21-26.
95. Wang K, Amdur RJ, Mendenhall WM, et al. Impact of post-chemoradiotherapy superselective/selective neck dissection on patient reported quality of life. *Oral Oncol.* 2016;58:21-26.
96. Spalthoff S, Zimmerer R, Jehn P, Gellrich NC, Handschel J, Kruskemper G. Neck Dissection's Burden on the Patient: Functional and Psychosocial Aspects in 1,652 Patients With Oral Squamous Cell Carcinomas. *J Oral Maxillofac Surg.* 2017;75(4):839-849.
97. Cappiello J, Piazza C, Giudice M, De Maria G, Nicolai P. Shoulder disability after different selective neck dissections (levels II-IV versus levels II-V): a comparative study. *Laryngoscope.* 2005;115(2):259-263.
98. Giordano L, Sarandria D, Fabiano B, Del Carro U, Bussi M. Shoulder function after selective and superselective neck dissections: clinical and functional outcomes. *Acta Otorhinolaryngol Ital.* 2012;32(6):376-379.
99. Saunders WH, Johnson EW. Rehabilitation of the shoulder after radical neck dissection. *Ann Otol Rhinol Laryngol.* 1975;84(6):812-816.
100. Takimoto T, Ishikawa S, Tanaka S, Masuda K, Umeda R. Development of significant sternoclavicular joint hypertrophy following radical neck dissection. *ORL J Otorhinolaryngol Relat Spec.* 1989;51(5):317-320.
101. Patten C, Hillel AD. The 11th nerve syndrome. Accessory nerve palsy or adhesive capsulitis? *Arch Otolaryngol Head Neck Surg.* 1993;119(2):215-220.
102. van den Berg MG, Rasmussen-Conrad EL, Gwasara GM, Krabbe PF, Naber AH, Merkx MA. A prospective study on weight loss and energy intake in patients with head and neck cancer, during diagnosis, treatment and revalidation. *Clin Nutr.* 2006;25(5):765-772.
103. Silver HJ, Dietrich MS, Murphy BA. Changes in body mass, energy balance, physical function, and inflammatory state in patients with locally advanced head and neck cancer treated with concurrent chemoradiation after low-dose induction chemotherapy. *Head Neck.* 2007;29(10):893-900.
104. Rogers LQ, Courneya KS, Robbins KT, et al. Physical activity and quality of life in head and neck cancer survivors. *Supportive Care in Cancer.* 2006;14(10):1012-1019.
105. Rogers LQ, Courneya KS, Robbins KT, et al. Physical activity correlates and barriers in head and neck cancer patients. *Supportive Care in Cancer.* 2008;16(1):19-27.

106. McGarvey AC, Osmotherly PG, Hoffman GR, Chiarelli PE. Impact of Neck Dissection on Scapular Muscle Function: A Case-Controlled Electromyographic Study. *Arch Phys Med Rehabil.* 2013;94(1):113-119.
107. Short SO, Kaplan JN, Laramore GE, Cummings CW. Shoulder pain and function after neck dissection with or without preservation of the spinal accessory nerve. *Am J Surg.* 1984;148(4):478-482.
108. van Wilgen CP, Dijkstra PU, van der Laan BFAM, Plukker JT, Roodenburg JLN. Shoulder and neck morbidity in quality of life after surgery for head and neck cancer. *Head Neck.* 2004;26(10):839-844.
109. Swisher AK, Altaha R, Arbaugh J, et al. Neck and Shoulder Impairments and the Relationship to Quality of Life in Head and Neck Cancer Survivors. *Rehabil Oncol.* 2012;30(2):3-7.
110. Nibu K, Ebihara Y, Ebihara M, et al. Quality of life after neck dissection: a multicenter longitudinal study by the Japanese Clinical Study Group on Standardization of Treatment for Lymph Node Metastasis of Head and Neck Cancer. *Int J Clin Oncol.* 2010;15(1):33-38.
111. Aasheim T, Finsen V. The DASH and the QuickDASH instruments. Normative values in the general population in Norway. *J Hand Surg Eur Vol.* 2013.
112. Guru K, Manoor UK, Supe SS. A comprehensive review of head and neck cancer rehabilitation: physical therapy perspectives. *Indian J Palliat Care.* 2012;18(2):87-97.
113. Lauchlan DT, McCaul JA, McCarron T. Neck dissection and the clinical appearance of post-operative shoulder disability: the post-operative role of physiotherapy. *Eur J Cancer Care (Engl).* 2008;17(6):542-548.
114. McNeely ML, Parliament M, Courneya KS, et al. A pilot study of a randomized controlled trial to evaluate the effects of progressive resistance exercise training on shoulder dysfunction caused by spinal accessory neurapraxia/neurectomy in head and neck cancer survivors. *Head Neck.* 2004;26(6):518-530.
115. McNeely ML, Parliament MB, Seikaly H, et al. Effect of exercise on upper extremity pain and dysfunction in head and neck cancer survivors. *Cancer.* 2008;113(1):214-222.
116. Lauchlan DT, McCaul JA, McCarron T, Patil S, McManners J, McGarva J. An exploratory trial of preventative rehabilitation on shoulder disability and quality of life in patients following neck dissection surgery. *Eur J Cancer Care (Engl).* 2011;20(1):113-122.
117. McGarvey AC, Chiarelli PE, Osmotherly PG, Hoffman GR. Physiotherapy for accessory nerve shoulder dysfunction following neck dissection surgery: a literature review. *Head Neck.* 2011;33(2):274-280.
118. Straus SE, Richardson WS, Glasziou P, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM.* 3rd ed. New York: Elsevier Churchill Livingstone; 2005.
119. Villanueva R. Orthosis to correct shoulder pain and deformity after trapezius palsy. *Arch Phys Med Rehabil.* 1977;58(1):30-34.
120. Johnson EW, Aseff JN, Saunders W. Physical treatment of pain and weakness following radical neck dissection. *Ohio State Med J.* 1978;74(11):711-714.
121. Fialka V, Vinzenz K. Investigations into shoulder function after radical neck dissection. *J Craniomaxillofac Surg.* 1988;16(3):143-147.

122. Kizilay A, Kalcioglu MT, Saydam L, Ersoy Y. A new shoulder orthosis for paralysis of the trapezius muscle after radical neck dissection: a preliminary report. *Eur Arch Otorhinolaryngol.* 2006;263(5):477-480.
123. Salerno G, Cavaliere M, Foglia A, et al. The 11th Nerve Syndrome in Functional Neck Dissection. *Laryngoscope.* 2002;112(7):1299-1307.
124. Shimada Y, Chida S, Matsunaga T, Sato M, Hatakeyama K, Itoi E. Clinical results of rehabilitation for accessory nerve palsy after radical neck dissection. *Acta Otolaryngol.* 2007;127(5):491-497.
125. Chida S, Shimada Y, Matsunaga T, Sato M, Hatakeyama K, Mizoi K. Occupational therapy for accessory nerve palsy after radical neck dissection. *Tohoku J Exp Med.* 2002;196(3):157-165.
126. McNeely ML, Parliament MB, Seikaly H, et al. Sustainability of outcomes after a randomized crossover trial of resistance exercise for shoulder dysfunction in survivors of head and neck cancer. *Physiother Can.* 2015;67(1):85-93.
127. Carvalho A, Vital F, Soares B. Exercise interventions for shoulder dysfunction ini patients treated for head and neck cancer. *Cochrane Database Syst Rev.* 2012;4.
128. McGarvey AC, Hoffman GR, Osmotherly PG, Chiarelli PE. Maximizing shoulder function after accessory nerve injury and neck dissection surgery: A multicenter randomized controlled trial. *Head Neck.* 2015;37:1022-1031.
129. Gallagher KK, Sacco AG, Lee JS, et al. Association Between Multimodality Neck Treatment and Work and Leisure Impairment: A Disease-Specific Measure to Assess Both Impairment and Rehabilitation After Neck Dissection. *JAMA Otolaryngol Head Neck Surg.* 2015;141(10):888-893.
130. Deganello A, Battat N, Muratori E, et al. Acupuncture in shoulder pain and functional impairment after neck dissection: A prospective randomized pilot study. *Laryngoscope.* 2016;126(8):1790-1795.
131. Fong SS, Ng SS, Lee HW, et al. The effects of a 6-month Tai Chi Qigong training program on temporomandibular, cervical, and shoulder joint mobility and sleep problems in nasopharyngeal cancer survivors. *Integr Cancer Ther.* 2015;14(1):16-25.
132. Bigliani LU, Perez-Sanz JR, Wolfe IN. Treatment of trapezius paralysis. *J Bone Joint Surg Am.* 1985;67(6):871-877.
133. Olarte M AD. Accessory Nerve Palsy. *J Neurol Neurosurg Psychiatry.* 1977;40:1113-1116.
134. Clinton S KE, Pariser G, Nuss D. Physical therapy management of a maunal laborer following a modified radical neck dissection. *Rehabil Oncol.* 2007;25(2):3-9.
135. McNeely ML, Parliament MB, Seikaly H, et al. Predictors of adherence to an exercise program for shoulder pain and dysfunction in head and neck cancer survivors. *Supportive Care in Cancer.* 2012;20(3):515-522.
136. Rogers LQ, Anton PM, Fogleman A, et al. Pilot, randomized trial of resistance exercise during radiation therapy for head and neck cancer. *Head Neck.* 2012:n/a-n/a.
137. Jeannie Kozempel PDM. Benefits of Preoperative Assessment for Neck Dissection. *Rehabil Oncol.* 2009;27:16-18.
138. Richardson M, Grobert M, Meyer K. Randomized Controlled Trials 3: Measurement and Analysis of Patient-Reported Outcomes. In: Parfrey PS, Barrett BJ, eds. *J Clin Epidemiol.* Vol 1281. Springer New York; 2015:191-206.

139. McDowell I. *Measuring health: a guide to rating scales and questionnaires.* 3 ed. New York, NY: Oxford University Press; 2006.
140. U.S. Food and Drug Administration. *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims.* Washington, DC: U.S. Department of Health and Human Services; 2009.
141. Raykov T, Marcoulides GA. Measurement, measuring instruments, and psychometric theory. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:1-11.
142. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. *Med Educ.* 2010;44(1):109-117.
143. DeVellis RF. Classical Test Theory. *Medical Care.* 2006;44(11, Suppl 3):S50-S59.
144. Streiner D, Norman G. Health measurement scales: A practical guide to their development and use. *Oxford: Oxford University Press.* 1995:64.
145. Hambleton RK, Jones RW. Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice.* 1993:38-47.
146. Hambleton RK, Swaminathan H. Assumptions of Item Response Theory. In: *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing; 2010:15-31.
147. Raykov T, Marcoulides GA. Classical Test Theory. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:115-136.
148. Raykov T, Marcoulides GA. Reliability. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:137-146.
149. Michener LA, Leggin BG. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. *J Hand Ther.* 2001;14(2):68-76.
150. Raykov T, Marcoulides GA. Procedures for Estimating Reliability. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:147-181.
151. Smith EV, Jr. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *J Appl Meas.* 2001;2(3):281-311.
152. Raykov T, Marcoulides GA. Validity. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:183-222.
153. Baghaei P. The Rasch Model as a Construct Validation Tool. *Rasch Measure Trans.* 2008;22(2):1145-1146.
154. Hambleton RK, Swaminathan H. Some Background to Item Response Theory. In: *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing; 2010:1-14.
155. Raykov T, Marcoulides GA. Introduction to Item Response Theory. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:247-267.
156. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil.* 1989;70(12):857-860.
157. Hays RD, Morales LS, Reise SP. Item Response Theory and Health Outcomes Measurement in the 21st Centry. *Med Care.* 2000;38(9 Suppl):1128-1142.
158. Anastasi A, Urbina S. item Analysis. In: *Psychological Testing.* 11th ed. Upper Saddle River, New Jersey: Prentice Hall; 1997:172-202.
159. Boone WJ, Staver JR, Yale MS. Rating Scale Surveys. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:21-46.

160. Hambleton RK, Swaminathan H. The Information Function and its Applications. In: *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing; 2010:101-124.
161. Raykov T, Marcoulides GA. Fundamentals and Models of Item Response Theory. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:269-304.
162. CTB/McGraw Hill. Accuracy of Test Scores: Why IRT Models Matter. 2008; https://www.ctb.com. Accessed August 12, 2015.
163. Furr RM, Bacharach VR. Item Response Theory and Rasch Models. In: *Psychometrics: An Introduction.* Los Angeles: SAGE Publications; 2007:314-332.
164. Hambleton RK, Swaminathan H. Item Response Models. In: *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing; 2010:33-52.
165. Boone WJ, Staver JR, Yale MS. The Rasch Model and Item Response Theory Models: Identical, Similar, or Unique? In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:449-458.
166. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum.* 2007;57(8):1358-1362.
167. Boone WJ, Staver JR, Yale MS. Quality of Measurement and Sample Size. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:357-376.
168. Straube D, Campbell SK. Rater Discrimination Using the Visual Analog Scale of the Physical Therapist Clinical Performance Instrument. *J Phys Ther Educ.* 2003;17(1):33-38.
169. *Winsteps Rasch measurement* [computer program]. Beaverton, Orgeon: www.winsteps.com. 2014.
170. RUMM Laboratory. RUMM2030 [computer program]. 2010; http://www.rummlab.com.au/rumm2020/rumm2020.html.
171. *ConQuest: Multi-Aspect Test Software* [computer program]. Camberwell: Australian Council for Educational Research1997.
172. Linacre JM. Dimensionality: contrasts & variances. www.winsteps.com. Accessed May 15, 2018.
173. Linacre J. A user's guide to Winsteps Ministeps Rasch-model computer programs [version 3.74.0]. 2012; http://www.winsteps.com/index.htm.
174. Raykov T, Marcoulides GA. An Introduction to Factor Analysis. In: *Introduction to Psychometric Theory.* New York: Routledge Taylor & Francis Group; 2011:37-59.
175. Boone WJ, Staver JR, Yale MS. Understanding Person Measures. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:69-92.
176. Boone WJ, Staver JR, Yale MS. Wright Maps: First Steps. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:111-136.
177. Boone WJ, Staver JR, Yale MS. Fit. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:159-189.
178. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice.* 3rd ed. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall; 2009.
179. Hambleton RK, Swaminathan H. Ability Scales. In: *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing; 2010:53-72.

180. Boone WJ, Staver JR, Yale MS. Differential item Functioning. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:273-297.
181. Linacre JM. DIF-DPF-bias-interactions conceps. https://www.winsteps.com/winman/difconcepts.htm. Accessed July 10, 2018.
182. Linacre JM. Optimizing Rating Scale Category Effectiveness. *J Allied Measure.* 2002;3(1):85-106.
183. Boone WJ, Staver JR, Yale MS. How Well Does That Rating Scale Work? How Do You Know, Too? In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:191-216.
184. Wright BD. "Logits"? *Rasch Measure Trans.* 1993;7(2):288.
185. Boone WJ, Staver JR, Yale MS. Person Reliability, Item Reliability, and More. In: *Rasch Analysis in the Human Sciences.* New York: Springer; 2014:217-234.
186. Linacre J. Reliability and separation measures. http://www.winsteps.com/winman/reliability.htm. Accessed April 30, 2014.
187. Rothstein JM, Echternach JL. *Primer on Measurement: An Introductory Guide to Measurement Issues.* American Physical Therapy Association; 1993.
188. Messick S. Validity of Psychological Assessment. Validation of Inferences From Persons' Responses and Performances as Scientfic Inquiry Into Score Meaning. *Am Psychol.* 1995;50(9):741-749.
189. Messick S. Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher.* 1989;18(2):5-11.
190. Ringash J, Bernstein L, Cella D, et al. Outcomes Toolbox for Head and Neck Cancer Research. *Head Neck.* 2015;37:425-439.
191. Field-Fote E. Towards Optimal Practice. What can we gain from assessment of patient progress with standardized outcome measures? http://www.ptresearch.org/article/14/edge-taskforce. Accessed August 3, 2013.
192. Clark JR, Vesely M, Gilbert R. Scapular angle osteomyogenous flap in postmaxillectomy reconstruction: Defect, reconstruction, shoulder function, and harvest technique. *Head Neck.* 2008;30(1):10-20.
193. Koh CE, Morrison WA. Functional impairment after latissimus dorsi flap. *ANZ J Surg.* 2009;79(1-2):42-47.
194. Marchese C, Cristalli G, Pichi B, et al. Italian cross-cultural adaptation and validation of three different scales for the evaluation of shoulder pain and dysfunction after neck dissection: University of California - Los Angeles (UCLA) Shoulder Scale, Shoulder Pain and Disability Index (SPADI) and Simple Shoulder Test (SST). *Acta Otorhinolaryngol Ital.* 2012;32(1):12-17.
195. Baldwin ER, Baldwin TD, Lancaster JS, McNeely ML, Collins DF. Neuromuscular electrical stimulation and exercise for reducing trapezius muscle dysfunction in survivors of head and neck cancer: a case-series report. *Physiother Can.* 2012;64(3):317-324.
196. Goldstein DP, Ringash J, Irish JC, et al. Assessment of the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire for use in patients following neck dissection for head and neck cancer. *Head Neck.* 2015;37(2):234-242.
197. Nkenke E, Vairaktaris E, Stelzle F, Neukam FW, Stockmann P, Linke R. Osteocutaneous free flap including medial and lateral scapular crests: technical aspects, viability, and donor site morbidity. *J Reconstr Microsurg.* 2009;25(9):545-553.

198. Miles BA, Gilbert RW. Maxillary reconstruction with the scapular angle osteomyogenous free flap. *Arch Otolaryngol Head Neck Surg.* 2011;137(11):1130-1135.
199. Ghiam MK, Mannion K, Dietrich MS, Stevens KL, Gilbert J, Murphy BA. Assessment of musculoskeletal impairment in head and neck cancer patients. *Supportive Care in Cancer.* 2017;25(7):2085-2092.
200. Kirkley A, Griffin S, Dainty K. Scoring Systems for the Functional Assessment of the Shoulder. *Arthoscopy.* 2003;19(10):1109-1120.
201. Richards RR, An K-N, Bigliani LU, et al. A standardized method for the assessment of shoulder function. *J Shoulder Elbow Surg.* 1994;3(6):347-352.
202. Celik D, Atalar AC, Demirhan M, Dirican A. Translation, cultural adaptation, validity and reliability of the Turkish ASES questionnaire. *Knee Surg Sports Traumatol Arthrosc.* 2013;21:2184-2189.
203. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. *J Shoulder Elbow Surg.* 2002;11(6):587-594.
204. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. *Ann Rheum Dis.* 2004;63(4):335-341.
205. Angst F, Schwyzer H-K, Aeschlimann A, Simmen BR, Goldhahn J. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and Its Short Version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society Standardized Shoulder Assessment Form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). *Arthritis Care Res.* 2011;63(S11):S174-S188.
206. Goldhahn J, Angst F, Drerup S, Pap G, Simmen BR, Mannion AF. Lessons learned during the cross-cultural adaptation of the American Shoulder and Elbow Surgeons shoulder form into German. *J Shoulder Elbow Surg.* 2008;17(2):248-254.
207. Padua R, Padua L, Ceccarelli E, Bondi R, Alviti F, Castagna A. Italian version of ASES questionnaire for shoulder assessment: cross-cultural adaptation and validation. *Musculoskelet Surg.* 2010;94 Suppl 1:S85-90.
208. Yahia A, Guermazi M, Khmekhem M, Ghroubi S, Ayedi K, Elleuch MH. Translation into Arabic and validation of the ASES index in assessment of shoulder disabilities. *Ann Phys Rehabil Med.* 2011;54(2):59-72.
209. Piitulainen K, Paloneva J, Ylinen J, Kautiainen H, Hakkinen A. Reliability and validity of the Finnish version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section. *BMC Musculoskelet Disord.* 2014;15:272.
210. Vrotsou K, Cuellar R, Silio F, et al. Patient self-report section of the ASES questionnaire: a Spanish validation study using classical test theory and the Rasch model. *Health Qual Life Outcomes.* 2016;14(1):147.
211. Knaut LA, Moser AD, Melo Sde A, Richards RR. Translation and cultural adaptation to the portuguese language of the American Shoulder and Elbow Surgeons Standardized Shoulder assessment form (ASES) for evaluation of shoulder function. *Rev Bras Reumatol.* 2010;50(2):176-189.

212. Moser AD, Knaut LA, Zotz TG, Scharan KO. Validity and reliability of the Portuguese version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form. *Rev Bras Reumatol.* 2012;52(3):348-356.
213. Razmjou H, Bean A, van Osnabrugge V, MacDermid JC, Holtby R. Cross-sectional and longitudinal construct validity of two rotator cuff disease-specific outcome measures. *BMC Musculoskelet Disord.* 2006;7:26.
214. Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. *J Bone Joint Surg Am.* 2005;87(9):2006-2011.
215. Cook C, Hegedus E, Goode A, Mina C, Pietrobon R, Higgins LD. Relative validity of the modified American Shoulder and Elbow Surgeons (M-ASES) questionnaire using item response theory. *Rheumatology International.* 2008;28(3):217-223.
216. Angst F, Pap G, Mannion AF, et al. Comprehensive assessment of clinical outcome and quality of life after total shoulder arthroplasty: Usefulness and validity of subjective outcome measures. *Arthritis Care Res.* 2004;51(5):819-828.
217. Angst F, Goldhahn J, Drerup S, Aeschlimann A, Schwyzer H-K, Simmen BR. Responsiveness of six outcome assessment instruments in total shoulder arthroplasty. *Arthritis Care Res.* 2008;59(3):391-398.
218. Napora JK, Grimberg D, Childs BR, Vallier HA. Factors Affecting Functional Outcomes After Clavicle Fracture. *J Am Acad Orthop Surg.* 2016;24(10):721-727.
219. Sallay PI, Reed L. The measurement of normative American Shoulder and Elbow Surgeons scores. *J Shoulder Elbow Surg.* 2003;12(6):622-627.
220. Clarke MG, Dewing CB, Schroder DT, Solomon DJ, Provencher MT. Normal shoulder outcome score values in the young, active adult. *J Shoulder Elbow Surg.* 2009;18(3):424-428.
221. Beaton D, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. *J Shoulder Elbow Surg.* 1998;7(6):565-572.
222. Cook KF, Roddey TS, Olson SL, Gartsman GM, Valenzuela FF, Hanten WP. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. *J Orthop Sports Phys Ther.* 2002;32(7):336-346.
223. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am.* 2010;92(2):296-303.
224. Macdermid JC, Khadilkar L, Birmingham TB, Athwal GS. Validity of the QuickDASH in patients with shoulder-related disorders undergoing surgery. *J Orthop Sports Phys Ther.* 2015;45(1):25-36.
225. Beckmann JT, Hung M, Bounsanga J, Wylie JD, Granger EK, Tashjian RZ. Psychometric evaluation of the PROMIS Physical Function Computerized Adaptive Test in comparison to the American Shoulder and Elbow Surgeons score and Simple Shoulder Test in patients with rotator cuff disease. *J Shoulder Elbow Surg.* 2015;24(12):1961-1967.
226. Skutek M, Fremerey RW, Zeichen J, Bosch U. Outcome analysis following open rotator cuff repair. Early effectiveness validated using four different shoulder assessment scales. *Arch Orthop Trauma Surg.* 2000;120(7-8):432-436.

227. Leggin BG, Michener LA, Shaffer MA, Brenneman SK, Iannotti JP, Williams GR, Jr. The Penn shoulder score: reliability and validity. *J Orthop Sports Phys Ther.* 2006;36(3):138-151.
228. Wright RW, Baumgarten KM. Shoulder Outcome Measures. *J Am Acad Orthop Surg.* 2010;18:436-444.
229. Schmidt S, Ferrer M, Gonzalez M, et al. Evaluation of shoulder-specific patient-reported outcome measures: a systematic and standardized comparison of available evidence. *J Shoulder Elbow Surg.* 2014;23(3):434-444.
230. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. *Clin Orthop Relat Res.* 1987(214):160-164.
231. Constant CR, Gerber C, Emery RJ, Sojbjerg JO, Gohlke F, Boileau P. A review of the Constant score: modifications and guidelines for its use. *J Shoulder Elbow Surg.* 2008;17(2):355-361.
232. Ban I, Troelsen A, Christiansen DH, Svendsen SW, Kristensen MT. Standardised test protocol (Constant Score) for evaluation of functionality in patients with shoulder disorders. *Dan Med J.* 2013;60(4):A4608.
233. Herr MW, Bonanno A, Montalbano LA, Deschler DG, Emerick KS. Shoulder function following reconstruction with the supraclavicular artery island flap. *Laryngoscope.* 2014;124(11):2478-2483.
234. Yao M, Yang L, Cao ZY, et al. Translation and cross-cultural adaptation of the Shoulder Pain and Disability Index (SPADI) into Chinese. *Clinical Rheumatology.* 2017;36(6):1419-1426.
235. Slobogean GP, Slobogean BL. Measuring shoulder injury function: Common scales and checklists. *Injury.* 2011;42(3):248-252.
236. Akgun K, Aktas I, Uluc K. Conservative treatment for late-diagnosed spinal accessory nerve injury. *Am J Phys Med Rehabil.* 2008;87(12):1015-1021.
237. Chepeha DB, Khariwala SS, Chanowski EJ, et al. Thoracodorsal artery scapular tip autogenous transplant: vascularized bone with a long pedicle and flexible soft tissue. *Arch Otolaryngol Head Neck Surg.* 2010;136(10):958-964.
238. Chepeha DB, Taylor RJ, Chepeha JC, et al. Functional assessment using Constant's Shoulder Scale after modified radical and selective neck dissection. *Head Neck.* 2002;24(5):432-436.
239. Murer K, Huber GF, Haile SR, Stoeckli SJ. Comparison of morbidity between sentinel node biopsy and elective neck dissection for treatment of the n0 neck in patients with oral squamous cell carcinoma. *Head Neck.* 2011;33(9):1260-1264.
240. Pfister DG, Cassileth BR, Deng GE, et al. Acupuncture for pain and dysfunction after neck dissection: results of a randomized controlled trial. *J Clin Oncol.* 2010;28(15):2565-2570.
241. Schiefke F, Akdemir M, Weber A, Akdemir D, Singer S, Frerich B. Function, postoperative morbidity, and quality of life after cervical sentinel node biopsy and after selective neck dissection. *Head Neck.* 2009;31(4):503-512.
242. Hudak P, Amadio P, Bombardier C, The Upper Extremity Collaborative Group. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. *Am J Ind Med.* 1996;29:602 - 608.
243. Kennedy CA, Beaton D, Soloway S, McConnell S, Bombardier C. *The DASH ouctome measure user's manual.* 3 ed. Toronto: Institute for Work & Health; 2011.

244. The Disability of the Arm, Shoulder and Hand (DASH) Outcome Measure. http://www.dash.iwh.on.ca/home. Accessed September 23, 2018.
245. Beaton DE, Wright JG, Katz JN. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am.* 2005;87(5):1038-1046.
246. Gabel CP, Yelland M, Melloh M, Burkett B. A modified QuickDASH-9 provides a valid outcome instrument for upper limb function. *BMC Musculoskelet Disord.* 2009;10:161.
247. Kennedy CA, Beaton DE. A user's survey of the clinical application and content validity of the DASH (Disabilities of the Arm, Shoulder and Hand) outcome measure. *J Hand Ther.* 2017;30(1):30-40 e32.
248. Hsu JE, Nacke E, Park MJ, Sennett BJ, Huffman GR. The Disabilities of the Arm, Shoulder, and Hand questionnaire in intercollegiate athletes: Validity limited by ceiling effect. *J Shoulder Elbow Surg.* 2010;19(3):349-354.
249. Hunsaker FG, Cioffi DA, Amadio PC, Wright JG, Caughlin B. The American Academy of Orthopaedic Surgeons Outcomes Insturments Normative Values from the General Population. *J Bone Joint Surg Am.* 2002;84A(2):208-215.
250. Gummesson C, Atroshi I, Ekdahl C. The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery. *BMC Musculoskelet Disord.* 2003;4(1):11.
251. Bilberg A, Bremell T, Mannerkorpi K. Disability of the Arm, Shoulder and Hand Questionnaire in Swedish Patients with Rheumatoid Arthritis: A Validity Study. *Journal Rehabil Med.* 2012;44(1):7-11.
252. Raven EEJ, Haverkamp D, Siervelt IN, et al. Construct Validity and Reliability of the Disability ofArm, Shoulder and Hand Questionnaire for Upper Extremity Complaints in Rheumatoid Arthritis. *J Rheumatol.* 2008;35(12):2334-2338.
253. Navsarikar A, Gladman DD, Husted JA, Cook RJ. Validity assessment of the disabilities of arm, shoulder, and hand questionnaire (DASH) for patients with psoriatic arthritis. *J Rheumatol.* 1999;26(10):2191-2194.
254. Huisstede BM, Feleus A, Bierma-Zeinstra SM, Verhaar JA, Koes BW. Is the disability of the arm, shoulder, and hand questionnaire (DASH) also valid and responsive in patients with neck complaints. *Spine (Phila Pa 1976).* 2009;34(4):E130-138.
255. Mehta S, Macdermid JC, Carlesso LC, McPhee C. Concurrent validation of the DASH and the QuickDASH in comparison to neck-specific scales in patients with neck pain. *Spine (Phila Pa 1976).* 2010;35(24):2150-2156.
256. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol.* 2004;57(10):1008-1018.
257. Rysstad T, Roe Y, Haldorsen B, Svege I, Strand LI. Responsiveness and minimal important change of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire (DASH) in patients with subacromial pain syndrome. *BMC Musculoskelet Disord.* 2017;18(1):248.
258. Staples MP, Forbes A, Green S, Buchbinder R. Shoulder-specific disability measures showed acceptable construct validity and responsiveness. *J Clin Epidemiol.* 2010;63(2):163-170.
259. Slobogean GP, Noonan VK, O'Brien PJ. The reliability and validity of the Disabilities of Arm, Shoulder, and Hand, EuroQol-5D, Health Utilities Index, and Short Form-6D

outcome instruments in patients with proximal humeral fractures. *J Shoulder Elbow Surg.* 2010;19(3):342-348.
260. Novak CB. Usefulness of the Disabilities of the Arm, Shoulder and Hand (DASH) to Assess Patients with Peripheral Nerve Injury. *The DASH and QuickDASH Outcome Measures e-Bulletin.* Winter 2012. http://dash.iwh.on.ca/system/files/dash_e-bulletin_2012_winter.pdf&sa=U&ei=pTxjU_bcNIWo0QXj_IHoDQ&ved=0CDkQFjAG&usg=AFQjCNEfnZcpAJ6EQ15PVdmLdKZy_6Fc7g. Accessed March 21, 2015.
261. Davies C, Brockopp D, Moe K. Internal consistency of the Disability of the Arm, Shoulder and Hand (DASH) Outcome Measure in Asessing Functional Status Among Breast Cancer Survivors. *Rehabil Oncol.* 2013;31(4):6-12.
262. Smoot B, Wong J, Cooper B, et al. Upper extremity impairments in women with or without lymphedema following breast cancer treatment. *Journal of Cancer Survivorship.* 2010;4(2):167-178.
263. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *J Hand Ther.* 2001;14:128 - 146.
264. Franchignoni F, Vercelli S, Giordano A, Sartorio F, Bravini E, Ferriero G. Minimal clinically important difference of the disabilities of the arm, shoulder and hand outcome measure (DASH) and its shortened version (QuickDASH). *J Orthop Sports Phys Ther.* 2014;44(1):30-39.
265. Roy JS, MacDermid JC, Woodhouse LJ. Measuring shoulder function: a systematic review of four questionnaires. *Arthritis Rheum.* 2009;61(5):623-632.
266. Lehman LA, Sindhu BS, Shechtman O, Romero S, Velozo CA. A Comparison of the Ability of Two Upper Extremity Assessments to Measure Change in Function. *J Hand Ther.* 2010;23(1):31-40.
267. Beaton D, Katz J, Fossel A, Wright J, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *J Hand Ther.* 2001;14:128 - 146.
268. Membrilla-Mesa MD, Cuesta-Vargas AI, Pozuelo-Calvo R, Tejero-Fernandez V, Martin-Martin L, Arroyo-Morales M. Shoulder pain and disability index: cross cultural validation and evaluation of psychometric properties of the Spanish version. *Health Qual Life Outcomes.* 2015;13:200.
269. Gorter RR, Vos CG, Halmans J, Hartemink KJ, Paul MA, Oosterhuis JWA. Evaluation of arm function and quality of life after trimodality treatment for superior sulcus tumours. *Interact Cardiovasc Thorac Surg.* 2013;16(1):44-48.
270. Roerink SH, Coolen L, Schenning ME, et al. High prevalence of self-reported shoulder complaints after thyroid carcinoma surgery. *Head Neck.* 2017;39(2):260-268.
271. Fang QG, Shi S, Zhang X, Li ZN, Liu FY, Sun CF. Upper extremity morbidity after radial forearm flap harvest: a prospective study. *J Int Med Res.* 2014;42(1):231-235.
272. Gabel CP, Michener LA, Melloh M, Burkett B. Modification of the upper limb functional index to a three-point response improves clinimetric properties. *J Hand Ther.* 2010;23(1):41-51; quiz 52.

273. Polson K, Reid D, McNair PJ, Larmer P. Responsiveness, minimal importance difference and minimal detectable change scores of the shortened disability arm shoulder hand (QuickDASH) questionnaire. *Man Ther.* 2010;15(4):404-407.
274. Roy JS, MacDermid JC, Amick BC, 3rd, et al. Validity and responsiveness of presenteeism scales in chronic work-related upper-extremity disorders. *Phys Ther.* 2011;91(2):254-266.
275. Stover B, Silverstein B, Wickizer T, Martin DP, Kaufman J. Accuracy of a disability instrument to identify workers likely to develop upper extremity musculoskeletal disorders. *J Occup Rehabil.* 2007;17(2):227-245.
276. Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. *J Shoulder Elbow Surg.* 2009;18(6):920-926.
277. Fan ZJ, Smith CK, Silverstein BA. Responsiveness of the QuickDASH and SF-12 in workers with neck or upper extremity musculoskeletal disorders: one-year follow-up. *J Occup Rehabil.* 2011;21(2):234-243.
278. Fan ZJ, Smith CK, Silverstein BA. Assessing validity of the QuickDASH and SF-12 as surveillance tools among workers with neck or upper extremity musculoskeletal disorders. *J Hand Ther.* 2008;21(4):354-365.
279. Quatman-Yates CC, Gupta R, Paterno MV, Schmitt LC, Quatman CE, Ittenbach RF. Internal Consistency and Validity of the QuickDASH Instrument for Upper Extremity Injuries in Older Children. *J Pediatr Orthop.* 2013.
280. Wu A, Edgar DW, Wood FM. The QuickDASH is an appropriate tool for measuring the quality of recovery after upper limb burn injury. *Burns.* 2007;33(7):843-849.
281. Budd HR, Larson D, Chojnowski A, Shepstone L. The QuickDASH score: a patient-reported outcome measure for Dupuytren's surgery. *J Hand Ther.* 2011;24(1):15-20; quiz 21.
282. Niekel MC, Lindenhovius AL, Watson JB, Vranceanu AM, Ring D. Correlation of DASH and QuickDASH with measures of psychological distress. *J Hand Surg Am.* 2009;34(8):1499-1505.
283. Jerosch-Herold C, Chester R, Shepstone L. Rasch Model Analysis Gives New Insights Into the Structural Validity of the QuickDASH in Patients With Musculoskeletal Shoulder Pain. *J Orthop Sports Phys Ther.* 2017;47(9):664-672.
284. Leblanc M, Stineman M, Demichele A, Stricker C, Mao JJ. Validation of QuickDASH Outcome Measure in Breast Cancer Survivors for Upper Extremity Disability. *Arch Phys Med Rehabil.* 2014;95(3):493-498.
285. Angst F, Goldhahn J, Drerup S, Flury M, Schwyzer HK, Simmen BR. How sharp is the short QuickDASH? A refined content and validity analysis of the short form of the disabilities of the shoulder, arm and hand questionnaire in the strata of symptoms and function and specific joint conditions. *Qual Life Res.* 2009;18(8):1043-1051.
286. Kennedy CA, Beaton DE, Smith P, et al. Measurement properties of the QuickDASH (Disabilities of the Arm, Shoulder and Hand) outcome measure and cross-cultural adaptations of the QuickDASH: a systematic review. *Qual Life Res.* 2013.
287. Gummesson C, Ward MM, Atroshi I. The shortened disabilities of the arm, shoulder and hand questionnaire (QuickDASH): validity and reliability based on responses within the full-length DASH. *BMC Musculoskelet Disord.* 2006;7:44.

288. Chester R, Jerosch-Herold C, Lewis J, Shepstone L. The SPADI and QuickDASH Are Similarly Responsive in Patients Undergoing Physical Therapy for Shoulder Pain. *J Orthop Sports Phys Ther.* 2017;47(8):538-547.
289. Brown EN, Chaudhry A, Mithani SK, et al. Long-term vascular, motor, and sensory donor site outcomes after ulnar forearm flap harvest. *J Reconstr Microsurg.* 2014;30(2):115-120.
290. Rogers SN, Scott B, Lowe D. An evaluation of the shoulder domain of the University of Washington quality of life scale. *Br J Oral Maxillofac Surg.* 2007;45(1):5-10.
291. Lee J, Kwon IS, Bae EH, Chung WY. Comparative analysis of oncological outcomes and quality of life after robotic versus conventional open thyroidectomy with modified radical neck dissection in patients with papillary thyroid carcinoma and lateral neck node metastases. *J Clin Endocrinol Metab.* 2013;98(7):2701-2708.
292. Parikh S, Tedman BM, Scott B, Lowe D, Rogers SN. A double blind randomised trial of IIb or not IIb neck dissections on electromyography, clinical examination, and questionnaire-based outcomes: a feasibility study. *Br J Oral Maxillofac Surg.* 2012;50(5):394-403.
293. Croft P, Pope D, Zonca M, O'Neill T, Silman A. Measurement of shoulder related disability: results of a validation study. *Ann Rheum Dis.* 1994;53(8):525-528.
294. van der Windt DA, van der Heijden GJ, de Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. *Ann Rheum Dis.* 1998;57(2):82-87.
295. van der Heijden GJ, Leffers P, Bouter LM. Shoulder disability questionnaire design and responsiveness of a functional status measure. *J Clin Epidemiol.* 2000;53(1):29-38.
296. Alvarez-Nemegyei J, Puerto-Ceballos I, Guzman-Hau W, Bassol-Perea A, Nuno-Gutierrez BL. Development of a Spanish-language version of the Shoulder Disability Questionnaire. *J Clin Rheumatol.* 2005;11(4):185-187.
297. Ozsahin M, Akgun K, Aktas I, Kurtais Y. Adaptation of the Shoulder Disability Questionnaire to the Turkish population, its reliability and validity. *Int J Rehabil Res.* 2008;31(3):241-245.
298. Choi Y, Park JW, Noh S, Kim MS, Park YH, Sung DH. Reliability, Validity, and Responsiveness of the Korean Version of the Shoulder Disability Questionnaire and Shoulder Rating Questionnaire. *Ann Rehabil Med.* 2015;39(5):705-717.
299. de Winter AF, van der Heijden GJ, Scholten RJ, van der Windt DA, Bouter LM. The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study. *J Clin Epidemiol.* 2007;60(11):1156-1163.
300. Paul A, Lewis M, Shadforth MF, Croft PR, Van Der Windt DA, Hay EM. A comparison of four shoulder-specific questionnaires in primary care. *Ann Rheum Dis.* 2004;63(10):1293-1299.
301. Ojo B, Genden EM, Teng MS, Milbury K, Misiukiewicz KJ, Badr H. A systematic review of head and neck cancer quality of life assessment instruments. *Oral Oncol.* 2012;48(10):923-937.
302. Stuiver MM, van Wilgen CP, de Boer EM, et al. Impact of shoulder complaints after neck dissection on shoulder disability and quality of life. *Otolaryngol Head Neck Surg.* 2008;139(1):32-39.

303. van Wilgen CP, Dijkstra PU, van der Laan BF, Plukker JT, Roodenburg JL. Shoulder complaints after nerve sparing neck dissections. *Int J Oral Maxillofac Surg.* 2004;33(3):253-257.
304. Cho JG, Lee N, Park MW, et al. Measurement of the trapezius muscle volume: A new assessment strategy of shoulder dysfunction after neck dissection for the treatment of head and neck cancers. *Head Neck.* 2015;37(5):619-623.
305. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. *Arthritis Care Res.* 1991;4(4):143-149.
306. Breckenridge JD, McAuley JH. Shoulder Pain and Disability Index (SPADI). *J Physiother.* 2011;57(3):197.
307. Williams JW, Jr., Holleman DR, Jr., Simel DL. Measuring shoulder function with the Shoulder Pain and Disability Index. *J Rheumatol.* 1995;22(4):727-732.
308. Heald SL, Riddle DL, Lamb RL. The shoulder pain and disability index: the construct validity and responsiveness of a region-specific disability measure. *Phys Ther.* 1997;77(10):1079-1089.
309. Torres-Lacomba M, Sanchez-Sanchez B, Prieto-Gomez V, et al. Spanish cultural adaptation and validation of the shoulder pain and disability index, and the oxford shoulder score after breast cancer surgery. *Health Qual Life Outcomes.* 2015;13:63.
310. Angst F, Goldhahn J, Pap G, et al. Cross-cultural adaptation, reliability and validity of the German Shoulder Pain and Disability Index (SPADI). *Rheumatology (Oxford).* 2007;46(1):87-92.
311. Puga VO, Lopes AD, Costa LO. Assessment of cross-cultural adaptations and measurement properties of self-report outcome measures relevant to shoulder disability in Portuguese: a systematic review. *Rev Bras Fisioter.* 2012;16(2):85-93.
312. Thoomes-de Graaf M, Scholten-Peeters W, Duijn E, et al. The Responsiveness and Interpretability of the Shoulder Pain and Disability Index. *J Orthop Sports Phys Ther.* 2017;47(4):278-286.
313. Alsanawi HA, Alghadir A, Anwer S, Roach KE, Alawaji A. Cross-cultural adaptation and psychometric properties of an Arabic version of the Shoulder Pain and Disability Index. *Int J Rehabil Res.* 2015;38(3):270-275.
314. Roddey TS, Olson SL, Cook KF, Gartsman GM, Hanten W. Comparison of the University of California-Los Angeles Shoulder Scale and the Simple Shoulder Test with the shoulder pain and disability index: single-administration reliability and validity. *Phys Ther.* 2000;80(8):759-768.
315. Huang H, Grant JA, Miller BS, Mirza FM, Gagnier JJ. A Systematic Review of the Psychometric Properties of Patient-Reported Outcome Instruments for Use in Patients With Rotator Cuff Disease. *Am J Sports Med.* 2015;43(10):2572-2582.
316. van den Dungen IA, Verhagen CA, van der Graaf WT, van den Berg JP, Vissers KC, Engels Y. Feasibility and impact of a physical exercise program in patients with advanced cancer: a pilot study. *J Palliat Med.* 2014;17(10):1091-1098.
317. Lippitt SB, Harryman DT, Matsen FA. A practical tool for evaluation of function: the Simple Shoulder Test. In: Matsen FA, Fu FH, Hawkins RJ, eds. *The shoulder: a balance of mobility and stability.* Rosemont (IL): American Academy of Orthopedic Surgeons; 1993:545-559.
318. Beaton DE, Richards RR. Measuring function of the shoulder. A cross-sectional comparison of five questionnaires. *J Bone Joint Surg Am.* 1996;78(6):882-890.

319. van Kampen DA, van Beers LW, Scholtes VA, Terwee CB, Willems WJ. Validation of the Dutch version of the Simple Shoulder Test. *J Shoulder Elbow Surg.* 2012;21(6):808-814.
320. Neto JO, Gesser RL, Steglich V, et al. Validation of the Simple Shoulder Test in a Portuguese-Brazilian population. Is the latent variable structure and validation of the Simple Shoulder Test Stable across cultures? *PLoS One.* 2013;8(5):e62890.
321. Naghdi S, Nakhostin Ansari N, Rustaie N, et al. Simple shoulder test and Oxford Shoulder Score: Persian translation and cross-cultural validation. *Arch Orthop Trauma Surg.* 2015;135(12):1707-1718.
322. Ebrahimzadeh MH, Vahedi E, Baradaran A, et al. Psychometric Properties of the Persian Version of the Simple Shoulder Test (SST) Questionnaire. *Arch Bone Jt Surg.* 2016;4(4):387-392.
323. Membrilla-Mesa MD, Tejero-Fernandez V, Cuesta-Vargas AI, Arroyo-Morales M. Validation and reliability of a Spanish version of Simple Shoulder Test (SST-Sp). *Qual Life Res.* 2015;24(2):411-416.
324. Raman J, MacDermid JC, Walton D, Athwal GS. Rasch analysis indicates that the Simple Shoulder Test is robust, but minor item modifications and attention to gender differences should be considered. *J Hand Ther.* 2017;30(3):348-358.
325. Godfrey J, Hamman R, Lowenstein S, Briggs K, Kocher M. Reliability, validity, and responsiveness of the simple shoulder test: psychometric properties by age and injury type. *J Shoulder Elbow Surg.* 2007;16(3):260-267.
326. Hsu JE, Russ SM, Somerson JS, Tang A, Warme WJ, Matsen FA, 3rd. Is the Simple Shoulder Test a valid outcome instrument for shoulder arthroplasty? *J Shoulder Elbow Surg.* 2017.
327. Romeo AA, Mazzocca A, Hang DW, Shott S, Bach BR, Jr. Shoulder scoring scales for the evaluation of rotator cuff repair. *Clin Orthop Relat Res.* 2004(427):107-114.
328. Hassan SJ, Weymuller EA, Jr. Assessment of quality of life in head and neck cancer patients. *Head Neck.* 1993;15(6):485-496.
329. Rogers SN, Ahad SA, Murphy AP. A structured review and theme analysis of papers published on 'quality of life' in head and neck cancer: 2000–2005. *Oral Oncology.* 2007;43(9):843-868.
330. Rogers SN, Forgie S, Lowe D, Precious L, Haran S, Tschiesner U. Development of the International Classification of Functioning, Disability and Health as a brief head and neck cancer patient questionnaire. *Int J Oral Maxillofac Surg.* 2010;39(10):975-982.
331. Ringash J, Bezjak A. A structured review of quality of life instruments for head and neck cancer patients. *Head Neck.* 2001;23(3):201-213.
332. Deleyiannis FW, Weymuller EA, Jr., Coltrera MD. Quality of life of disease-free survivors of advanced (stage III or IV) oropharyngeal cancer. *Head Neck.* 1997;19(6):466-473.
333. Weymuller EA, Yueh B, Deleyiannis FW, Kuntz AL, Alsarraf R, Coltrera MD. Quality of life in patients with head and neck cancer: lessons learned from 549 prospectively evaluated patients. *Arch Otolaryngol Head Neck Surg.* 2000;126(3):329-335; discussion 335-326.
334. Weymuller EA, Jr., Alsarraf R, Yueh B, Deleyiannis FW, Coltrera MD. Analysis of the performance characteristics of the University of Washington Quality of Life instrument

and its modification (UW-QOL-R). *Arch Otolaryngol Head Neck Surg.* 2001;127(5):489-493.
335. Rogers SN, Gwanne S, Lowe D, Humphris G, Yueh B, Weymuller EA, Jr. The addition of mood and anxiety domains to the University of Washington quality of life scale. *Head Neck.* 2002;24(6):521-529.
336. Ghazali N, Lowe D, Rogers SN. Enhanced patient reported outcome measurement suitable for head and neck cancer follow-up clinics. *Head Neck Oncol.* 2012;4:32.
337. Rogers SN, Lowe D, Yueh B, Weymuller EA, Jr. The physical function and social-emotional function subscales of the University of Washington Quality of Life Questionnaire. *Arch Otolaryngol Head Neck Surg.* 2010;136(4):352-357.
338. Rogers SN, Lowe D, Brown JS, Vaughan ED. A comparison between the University of Washington Head and Neck Disease-Specific measure and the Medical Short Form 36, EORTC QOQ-C33 and EORTC Head and Neck 35. *Oral Oncol.* 1998;34(5):361-372.
339. de Andrade FP, Biazevic MG, Toporcov TN, Togni J, de Carvalho MB, Antunes JL. Discriminant validity of the University of Washington quality of life questionnaire in the Brazilian context. *Rev Bras Epidemiol.* 2012;15(4):781-789.
340. Lowe D, Rogers SN. University of Washington Quality of Life Questionnaire (UW-QOL v4): Guidance for scoring and presentation. 5/25/2012; http://www.headandneckcancer.co.uk/File.ashx?id=10285. Accessed March 1, 2015.
341. Merseyside Regional Head and Neck Cancer Centre. UW-QOL v4 Questionnaire - translated versions. http://www.headandneckcancer.co.uk/For+professionals/Quality+of+Life+(QOL)/UW-QOLv4+Translations.aspx. Accessed March 1, 2015.
342. Lee YH, Lai YH, Yueh B, et al. Validation of the University of Washington Quality of Life Chinese Version (UWQOL-C) for head and neck cancer patients in Taiwan. *J Formos Med Assoc.* 2017;116(4):249-256.
343. Rampling T, King H, Mais KL, et al. Quality of life measurement in the head and neck cancer radiotherapy clinic: is it feasible and worthwhile? *Clin Oncol (R Coll Radiol).* 2003;15(4):205-210.
344. Cardoso LR, Rizzo CC, de Oliveira CZ, Dos Santos CR, Carvalho AL. Myofascial pain syndrome after head and neck cancer treatment: Prevalence, risk factors, and influence on quality of life. *Head Neck.* 2014.
345. Rogers SN, Lowe D. Screening for dysfunction to promote multidisciplinary intervention by using the University of Washington Quality of Life Questionnaire. *Arch Otolaryngol Head Neck Surg.* 2009;135(4):369-375.
346. Thoomes-de Graaf M, Scholten-Peeters GG, Schellingerhout JM, et al. Evaluation of measurement properties of self-administered PROMs aimed at patients with non-specific shoulder pain and "activity limitations": a systematic review. *Qual Life Res.* 2016;25(9):2141-2160.
347. Harrington S, Michener LA, Kendig T, Miale S, George SZ. Patient-Reported Shoulder Outcome Measures Utilized in Breast Cancer Survivors: A Systematic Review. *Arch Phys Med Rehabil.* 2013.
348. Bafus BT, Hughes RE, Miller BS, Carpenter JE. Evaluation of utility in shoulder pathology: Correlating the American Shoulder and Elbow Surgeons and Constant scores to the EuroQoL. *World J Orthop.* 2012;3(3):20-24.

349. Smith MV, Calfee RP, Baumgarten KM, Brophy RH, Wright RW. Upper Extremity-Specific Measures of Disability and Outcomes in Orthopaedic Surgery. *J Bone Joint Surg Am.* 2012;94(3):277-285.
350. Roe Y, Soberg HL, Bautz-Holter E, Ostensjo S. A systematic review of measures of shoulder pain and functioning using the International classification of functioning, disability and health (ICF). *BMC Musculoskelet Disord.* 2013;14:73.
351. Rogers SN, Heseltine N, Flexen J, Winstanley HR, Cole-Hawkins H, Kanatas A. Structured review of papers reporting specific functions in patients with cancer of the head and neck: 2006 - 2013. *Br J Oral Maxillofac Surg.* 2016;54(6):e45-51.
352. Linacre JM. Sample size and item callibration stability. *Rasch Measure Trans.* 1994;7(4):328.
353. Goldstein DP, Eskander A, Chepeha DB, Ringash J, Irish J, Davis AM. Response rates for mailout survey-driven studies in patients with head and neck cancer. *Head Neck.* 2010;32(12):1585-1591.
354. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDcap) - A metadata-driven methodology and workflow process for providing translational research information support. *J Biomed Inform.* 2009;42(2):377-381.
355. Fisher W. Rating Scale Instrument Quality Criteria. 2018; https://www.rasch.org/rmt/rmt211m.htm. Accessed June 6, 2018.
356. Eden M, Cheng MS. Comparison of 4 Shoulder-Related Outcome Measures in Individuals with Head and Neck Cancer Using Rasch Analysis. Paper presented at: American Physical Therapy Association Combined Section's Meeting2017; San Antonio, Texas.

Appendices

Appendix 1. Letter of Approval from the Mayo Clinic IRB

Principal Investigator Notification:

From: Mayo Clinic IRB

To: Melissa Eden

Re: **IRB Application #:** 15-005266

Title: Reliability and Validity of Five Shoulder-Specific Patient Reported Outcome Measures for use in Patients with Head and Neck Cancer

IRBe Protocol Version: 0.01
IRBe Version Date: 8/4/2015 10:51 AM

IRB Approval Date: 9/15/2015
IRB Expiration Date: 9/14/2016

The above referenced application is approved by expedited review procedures (45 CFR 46.110, item 5, 7). This approval is valid for a period of one year. The Reviewer conducted a risk-benefit analysis, and determined the study constitutes minimal risk research. The Reviewer determined that this research satisfies the requirements of 45 CFR 46.111. This study has IRB approval to accrue 250 adult subjects. The Reviewer noted that oral consent with HIPAA authorization is appropriate for this study. The oral consent script and HIPAA form were reviewed and approved as written. The Reviewer approved waiver of the requirement for the Investigator to obtain a signed consent form in accordance with 45 CFR 46.117 as justified by the Investigator. The DASH, Shoulder Pain and Disability Index, Neck Dissection Impairment Index, and Demographics Questionnaires are approved.

The Reviewer accepts the appointment of Mayo Clinic IRB as the IRB of record for the relying organization, Nova Southeastern University, and noted receipt of the fully executed IRB Authorization Agreement.

AS THE PRINCIPAL INVESTIGATOR OF THIS PROJECT, YOU ARE RESPONSIBLE FOR THE FOLLOWING RELATING TO THIS STUDY.

1) When applicable, use only IRB approved materials which are located under the documents tab of the IRBe workspace. Materials include consent forms, HIPAA, questionnaires, contact letters, advertisements, etc.

2) Submission to the IRB of any modifications to approved research along with any supporting documents for review and approval prior to initiation of the changes.

3) Submission to the IRB of all Unanticipated Problems Involving Risks to Subjects or Others (UPIRTSO).

4) Compliance with Mayo Clinic Institutional Policies.

Mayo Clinic Institutional Reviewer

Appendix 2. Oral Consent Template

Mayo Clinic: Office for Human Research Protection
Oral Consent Script

Protocol Title: Reliability and Validity of Five Shoulder-Specific Patient Reported Outcome Measures for use in Patients with Head and Neck Cancer
IRB #: 15-005266
Principal Investigator: Melissa M. Eden PT, DPT, PhD(c), OCS

You are being asked to participate in a research study that will compare several questionnaires that are often used to measure shoulder pain, stiffness or weakness in patients following surgery for head and neck cancer. You are being asked to participate in this study because you have had a neck dissection surgery within the past 18 months and are currently experiencing some level of shoulder dysfunction.

If you do not have shoulder pain, stiffness or weakness, or if you have already completed this packet at another appointment, please do not complete these forms. Please return this packet to your medical provider now.

If you agree to participate you will be asked to complete 4 questionnaires. One of the questionnaires asks you to provide basic information about yourself. The other 3 questionnaires require you to answer questions about your shoulder symptoms and your ability to perform various activities. It is anticipated that these questionnaires will take you 30 minutes to complete. All information collected will be stored in a secure database in order to protect your confidentiality. You will not receive payment for your participation.

There are no known risks to you for taking part in this research study. This study may not directly benefit you, but it will benefit the medical providers who participate in your care and the care of future patients.

Please understand your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty. Specifically, your current or future medical care at the Mayo Clinic will not be jeopardized if you choose not to participate.

If you have any questions about this research study you can contact me, the primary investigator, at (480) 342-0450. If you have any concerns, complaints, or general questions about research or your rights as a participant, please contact the Mayo Institutional Review Board (IRB) to speak to someone independent of the research team at 507-266-4000 or toll free at 866-273-4681.

Appendix 3. HIPAA Authorization to Use and Disclose Protected Health Information

MAYO CLINIC

HIPAA Authorization to Use and Disclose Protected Health Information

Name and Clinic Number

Approval Date:
Not to be used after:

Study Title: Reliability and Validity of Five Shoulder-Specific Patient Reported Outcome Measures for use in Patients with Head and Neck Cancer

IRB#:

Principal Investigator: Melissa M. Eden PT, DPT, PhD(c), OCS and Colleagues

During this research, information about your health will be collected. Under Federal law called the Privacy Rule, health information is private. However, there are exceptions to this rule, and you should know who may be able to see, use and share your health information for research and why they may need to do so. Information about you and your health cannot be used in this research study without your written permission. If you sign this form, it will provide that permission. You will be given a copy of this form.

Health information may be collected about you from:

- Past, present and future medical records.
- Research procedures, including research office visits, tests, interviews and questionnaires.

This information will be used and/or given to others to:

- Do the research.
- Report the results.
- See if the research was done correctly.

If the results of this study are made public, information that identifies you will not be used.

Your health information may be used or shared with:

- Mayo Clinic research staff involved in this study.

Your health information may also be shared with:

- The Mayo Clinic Institutional Review Board that oversees the research.
- Researchers involved in this study at other institutions.
- Federal and State agencies (such as the Food and Drug Administration, the Department of Health and Human Services, the National Institutes of Health and other United States agencies) or government agencies in other countries that oversee or review research.
- The sponsor(s) of this study and the people or groups it hires to help perform this research.
- A group that oversees the data (study information) and safety of this research.

MAYO CLINIC

HIPAA Authorization to Use and Disclose Protected Health Information

Name and Clinic Number

Approval Date:
Not to be used after:

Protection of your health information after it has been shared with others:
Mayo Clinic asks anyone who receives your health information from us to protect your privacy; however, once your information is shared outside Mayo Clinic, we cannot promise that it will remain private and it may no longer be protected by the Privacy Rule.

Your Privacy Rights
You do not have to sign this form, but if you do not, you cannot take part in this research study. Your decision won't change the access to medical care or any other benefits you get at Mayo Clinic now or in the future.

If you cancel your permission to use or share your health information, your participation in this study will end and no more information about you will be collected; however, information already collected about you in the study may continue to be used.

You can cancel your permission to use or share your health information at any time by sending a letter to the address below:
Mayo Clinic
Office for Human Research Protection
ATTN: Notice of Revocation of Authorization
200 1st Street SW
Rochester, MN 55905

Alternatively, you may cancel your permission by emailing the Mayo Clinic Research Subject Advocate at: researchsubjectadvocate@mayo.edu.

Please be sure to include in your letter or email:

- The name of the Principal Investigator,
- The study IRB number and /or study name, and
- Your contact information.

Your permission lasts forever, unless you cancel it.

Your signature documents your permission to use your protected health information for this research.

	/ / : AM/PM	
Printed Name	Date	Time

_____
Signature

Appendix 4. Patient Contact Letter
(Date)

(Name)
(Street Address)
(City, State Zip)

RE: *(first name) (last name)*
MC#: (*mc #)*

Protocol Title: Reliability and Validity of Five Shoulder-Specific Patient Reported Outcome Measures for use in Patients with Head and Neck Cancer
IRB #: 15-005266
Principal Investigator: Melissa M. Eden PT, DPT, PhD(c), OCS

Dear (*Mr., Ms, or Mrs.)*

You are being asked to participate in a research study that will compare several questionnaires that are often used to measure shoulder pain, stiffness or weakness in patients following surgery for head and neck cancer. You are being asked to participate in this study because you have had a neck dissection surgery within the past 18 months and are currently experiencing some level of shoulder dysfunction.

If you do not have shoulder pain, stiffness or weakness, or if you have already completed this packet at another appointment, please do not complete these forms.

If you agree to participate you will be asked to complete 4 questionnaires. One of the questionnaires asks you to provide basic information about yourself. The other 3 questionnaires require you to answer questions about your shoulder symptoms and your ability to perform various activities. It is anticipated that these questionnaires will take you 30 minutes to complete. All information collected will be stored in a secure database in order to protect your confidentiality. You will not receive payment for your participation. We have enclosed the questionnaires for you to complete. If you would like to, you may fill it out and return in the enclosed stamped envelope.

There are no known risks to you for taking part in this research study. This study may not directly benefit you, but it will benefit the medical providers who participate in your care and the care of future patients.

Please understand your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty. Specifically, your current or future medical care at the Mayo Clinic will not be jeopardized if you choose not to participate.

If you decide to participate, please read and sign the Authorization to Use and Disclose Protected Health Information form and return it with the questionnaire. We are not allowed to use the answers without your signature on the Authorization to Use and Disclose Protected Health Information form. A copy will be available through your Mayo Clinic patient portal if you have a portal account. Otherwise, a copy is available upon your request by marking the appropriate box on the last page.

Contact me at (480) 342-0450 if you have any questions about:

- Study tests and procedures
- Withdrawing from the research study
- Materials you receive

If you prefer, you may write to me at the address given below:

Melissa Eden PT, DPT
Physical Medicine and Rehabilitation
5777 East Mayo Boulevard
Phoenix, Arizona 85054

Contact the Mayo Institutional Review Board (IRB) to speak to someone independent of the research team at 507-266-4000 or toll free at 866-273-4681 if you have questions about:

- Rights of a research participant
- Use of your Protected Health Information
- Stopping your authorization to use your Protected Health Information

Research-related questions not listed above, or any research-related complaints may also be addressed to me. If you prefer to speak with someone independent of the research team, you may contact the Mayo Institutional Review Board (IRB).

If you prefer to complete the survey over the phone, or if you do not wish to participate, please indicate on the next page and return this letter since it will make a follow-up telephone call unnecessary. Thank you very much for your time and consideration.

Sincerely,

_____
Melissa M. Eden PT, DPT
Primary Investigator

RE: *(first name) (last name)*
MC#: *(mc #)*

☐ I would prefer to complete the survey over the phone. I am **enclosing** the **Authorization to Use and Disclose Protected Health Information form** only. Please call me.

Your name: _____
Telephone number: (___)___-_____
Today's date: __/__/__
Best time to call: ☐ Morning ☐ Afternoon ☐ Evening
Best day(s) to call:_____

☐ I am requesting that a copy of the Authorization to Use and Disclose Protected Health Information form be mailed to me.

☐ I am **not** willing to participate in this research study.

Appendix 5. Demographics Questionnaire

Patient Label:

Demographics Questionnaire

Are you currently experiencing shoulder discomfort, limited motion (stiffness) and/or weakness as a result of your neck surgery? Yes No

If your answer is "NO" to the above question you **do not** need to complete the remainder of this packet. Please return the packet to your medical provider. Thank you.

Please answer the following questions about yourself:

1. Which shoulder gives you trouble: Right Left Both

2. What is your height and weight? Height:_____ Weight: _____

3. Are you right or left handed? Right Left Ambidextrous

4. What is your ethnicity?

_____African American
_____Asian American
_____White, non-hispanic
_____White, Hispanic
_____Middle Eastern
_____Other

5. Have you been treated or received exercises for your shoulder from any of the following: (check all that apply)

_____Physician
_____Nurse
_____Physical Therapist
_____Occupational Therapist
_____Massage Therapist
_____Chiropractor
_____Acupunctruist
_____Personal Trainer
_____None

6. Please check the one box that best describes your shoulder **over the past seven days**.

_____I have no problems with my shoulder.
_____My shoulder is stiff but it has not affected my activity or strength.
_____Pain or weakness in my shoulder has caused me to change my work.
_____I cannot work due to problems in my shoulder.

Appendix 6. Data Mining Form

Data Mining Form

(to be completed by study personnel)

Patient Label:	State of Residence: Treatment Site: MCR MCJ MCA

Date of surgery:

Neck Dissection Type:	right	
	left	
	bilateral	

Levels Dissected:

Status of CN XI:		Right	Left
	preserved		
	sacraficed		

Diagnosis:

Stage:

Additional Interventions Received:	radiotherapy	
	chemotherapy	
	combined modality	

Appendix 7. Letter of Award for Research Grant

Oncology Section American Physical Therapy Association — Celebrate Living

June 8, 2015

Melissa M. Eden, PT, DPT, PhD, OCS
[redacted]

Dear Dr. Eden:

Your recent application for the Oncology Section's Research Grant has been reviewed by the Section's Research Committee. Congratulations, the Committee has chosen your application to receive the $5,000 grant.

The $5,000 check will be made payable to the Mayo Clinic and sent to:

Mayo Clinic Arizona
PO Box 860334
Minneapolis, MN 55486-0334

Sincerely,

Tori Marchese, PT, PhD
Research Committee Member and Manager of Section Grants, Oncology Section, APTA

Shana Harrington, PT, PhD, MTC, SCS
Chair, Research Committee, Oncology Section, APTA

APTA
American Physical Therapy Association
The Science of Healing. The Art of Caring.

Appendix 8. DASH & QuickDASH Summary Table

		DASH Fit Statistics (Item)				QuickDASH Fit Statistics (Item)		Response Category Frequency				
DASH Item (*item stem*)	**Measure (Model S.E.)**	**Infit Statistic (z-score)**	**Outfit Statistic (z-score)**	**Quick DASH Item**	**Measure (Model S.E.)**	**Infit Statistic (z-score)**	**Outfit Statistic (z-score)**	**1**	**2**	**3**	**4**	**5**
1. Open a tight or new jar (*open jar*)	49.3 (1.0)	***1.46 (3.7)***	***1.59 (4.2)***	1	49.1 (1.0)	***1.42 (3.4)***	***1.49 (3.5)***	61	61	32	18	7
2. Write (*write*)	66.6 (1.5)	1.30 (1.6)	1.60 (1.6)					141	26	9	2	1
3. Turn a key (*turn key*)	61.4 (1.7)	1.03 (0.2)	0.93 (-0.1)					144	26	7	0	0
4. Prepare a meal (*prepare meal*)	64.2 (1.4)	0.81 (-1.4)	***0.58 (-2.4)***					111	49	3	2	1
5. Push open a heavy door (*open door*)	52.9 (1.1)	0.94 (0.51)	0.98 (-0.1)					67	67	30	12	4
6. Place an object on a shelf above your head (*object overhead*)	40.0 (1.0)	1.10 (1.0)	1.05 (0.5)					20	59	56	32	12
7. Do heavy household chores (*heavy chores*)	42.2 (0.9)	1.02 (0.3)	0.99 (0.0)	2	41.7 (1.0)	1.13 (1.3)	1.18 (1.6)	36	50	50	24	16
8. Garden or do yard work (*garden/yard work*)	42.2 (1.0)	0.75 (-2.5)	0.74 (-2.5)					32	64	40	23	16
9. Make a bed (*make bed*)	56.7 (1.2)	0.84 (-1.3)	0.76 (-1.7)					83	59	28	3	3
10. Carry a shopping bag or briefcase (*carry bag*)	57.1 (1.2)	0.88 (-1.1)	0.84 (-1.2)	3	57.0 (1.2)	0.92 (-0.7)	0.88 (-0.9)	76	63	34	4	2
11. Carry a heavy object (over 10 lbs.) (*carry object*)	45.9 (1.0)	1.16 (1.4)	1.19 (1.7)					44	61	47	17	10
12. Change a light bulb overhead (*change bulb*)	40.0 (1.0)	0.91 (-0.8)	0.89 (-1.0)					26	58	41	32	17
13. Wash or blow dry your hair (*style hair*)	52.2 (1.0)	1.04 (0.4)	1.01 (0.1)					77	45	37	9	6
14. Wash your back (*wash back*)	44.2 (0.9)	0.87 (-1.2)	0.83 (-1.5)	4	43.8 (1.0)	0.93 (-0.5)	0.94 (-0.5)	44	61	34	23	13
15. Put on a pullover sweater (*don sweater*)	49.2 (1.1)	0.95 (-0.5)	0.95 (-0.4)					45	71	41	15	5
16. Use a knife to cut food (*cut food*)	66.5 (1.4)	0.89 (-0.8)	0.65 (-1.3)	5	66.5 (1.5)	0.91(-1.5)	0.61 (-1.5)	127	34	14	0	1
17. Recreational activities which require little effort (*recreation – light*)	60.7 (1.3)	0.98 (-0.1)	1.04 (0.2)					133	30	9	2	3
18. Recreational activities in which you take some force or impact through your arm, shoulder or hand (*recreation – force/impact*)	38.7 (1.0)	0.87 (-1.2)	0.83 (-1.6)	6	37.8 (1.0)	0.93 (-0.7)	0.88 (-1.1)	16	62	49	34	14
19. Recreational activities in which you move your arm freely (*recreation – free movement*)	39.7 (1.0)	0.94 (-0.5)	0.95 (-0.4)					29	43	51	33	17
20. Manage transportation needs (*manage transportation*)	59.0 (1.2)	1.11 (0.7)	0.94 (-.01)					122	34	12	2	4
21. Sexual activities (*sexual activities*)	53.1 (1.0)	***1.40 (2.5)***	***2.84 (4.7)***					109	30	13	7	10

22. During the past week, to what extent has your arm, shoulder, or hand problem interfered with your normal social activities? (*interfere - social*)	53.5 (1.0)	0.89 (-0.9)	0.87 (-0.8)	7	53.3 (1.0)	0.82 (-1.5)	0.72 (0.67)	83	54	28	10	5
23. During the past week, were you limited in your work or other regular daily activities as a result of your arm, shoulder or hand problem? (*limit activities*)	48.3 (1.0)	0.81 (-1.8)	0.86 (-1.3)	8	48.0 (1.0)	0.78 (-2.2)	0.86 (-1.3)	52	59	45	16	7
24. Arm, shoulder or hand pain (*pain*)	48.7 (1.1)	0.88 (-1.2)	0.90 (-0.9)	9	48.2 (1.1)	0.80 (-2.0)	0.82 (-1.8)	30	80	49	18	3
25. Arm, shoulder or hand pain when you performed any specific activity (*pain with activity*)	42.5 (1.1)	0.78 (-2.2)	0.77 (-2.4)					21	65	60	25	8
26. Tingling (pins and needles) in your arm, shoulder or hand (*tingling*)	54.4 (1.0)	1.36 (2.7)	***1.85 (3.7)***	10	54.3 (1.0)	1.24 (1.9)	***1.47 (2.2)***	97	41	25	12	5
27. Weakness in your arm, shoulder or hand (*weakness*)	38.7 (1.1)	0.94 (-0.6)	0.93 (-0.6)					9	66	69	26	9
28. Stiffness in your arm, shoulder or hand (*stiffness*)	40.8 (1.1)	0.97 (-0.3)	0.97 (-0.2)					13	66	64	28	7
29. During the past week, how much difficulty have you had sleeping because of the pain in your arm, shoulder or hand? (*sleep*)	50.6 (1.0)	1.17 (1.4)	***1.42 (2.8)***	11	50.3 (1.0)	1.04 (0.4)	1.04 (0.4)	69	55	37	11	7
30. I feel less capable, less confident or less useful because of my arm, shoulder or hand problem (*self-efficacy*)	40.5 (0.9)	1.23 (2.0)	1.17 (1.3)					41	32	33	60	14

Italicized text indicates item misfit, defined as Infit or Outfit statistic < 0.6 or > 1.4, z-score > 2.0

Appendix 9. SPADI Summary Table

	Full Scale			Subscales*			Response Category Frequency										
SPADI Item (*item stem*)	**Measure (Model S.E.)**	**Item Infit Statistic (z-score)**	**Item Outfit Statistic (z-score)**	**Measure (Model S.E.)**	**Item Infit Statistic (z-score)**	**Item Outfit Statistic (z-score)**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Pain (How severe is your pain?)																	
1. At its worst? (*pain-worst*)	43.9 (0.5)	1.08 (0.7)	1.19 (1.6)	43.3 (0.6)	0.89 (-0.9)	0.96 (-0.3)	16	13	24	29	15	24	8	11	23	7	8
2. When lying on the involved side? (*pain-involved side*)	48.4 (0.5)	1.20 (1.6)	1.27 (0.8)	51.2 (0.6)	0.88 (-1.0)	0.79 (-1.7)	36	20	31	20	10	13	13	10	14	6	5
3. Reaching for something on a high shelf? (*pain-shelf*)	44.4 (0.5)	0.74 (-2.5)	0.71 (-2.8)	50.8 (0.6)	1.14 (1.1)	1.06 (0.5)	14	20	28	28	11	20	12	13	12	11	8
4. Touching the back of your neck? (*pain-touch neck*)	50.1 (0.5)	0.85 (-1.2)	0.78 (-1.6)	51.2 (0.6)	0.95 (-0.3)	0.96 (-0.2)	42	22	28	27	12	12	11	10	6	1	6
5. Pushing with the involved arm? (*pain-pushing*)	49.8 (0.5)	1.02 (0.2)	0.88 (-1.0)	48.7 (0.6)	1.17 (1.3)	1.15 (1.2)	34	30	34	19	15	10	10	9	10	2	5
Disability (How much difficulty do you have?)																	
6. Washing your hair? (*disability-wash hair*)	53.6 (0.6)	0.99 (0.0)	0.85 (-1.0)	52.9 (0.6)	1.01 (0.1)	0.80 (-1.3)	65	30	22	17	10	13	5	7	3	1	4
7. Washing your back? (*disability-wash back*)	48.2 (0.5)	1.09 (0.8)	1.08 (0.6)	46.8 (0.5)	1.01 (0.2)	0.96 (-1.3)	36	29	26	22	10	12	7	10	9	5	10
8. Putting on an undershirt or jumper? (*disability–don shirt*)	51.6 (0.6)	0.69 (-2.9)	0.71 (-2.5)	50.6 (0.6)	0.65 (-3.2)	0.67 (-3.1)	36	30	31	25	15	14	4	12	7	0	3
9. Putting on a shirt that buttons	55.4 (0.6)	0.89 (-0.9)	0.88 (-0.7)	55.1 (0.6)	0.89 (-0.9)	0.81 (-1.1)	72	28	20	23	10	9	6	7	1	2	0

down the front? (*disability-buttons*)																	
10. Putting on your pants? (*disability-don pants*)	60.8 (0.7)	0.87 (-1.0)	0.70 (-1.6)	55.3 (0.7)	0.80 (-1.5)	0.66 (-1.7)	93	27	24	13	8	4	5	3	0	1	0
11. Placing an object on a high shelf? (*disability-shelf*)	43.2 (0.5)	1.19 (1.7)	1.22 (1.7)	40.7 (0.6)	1.23 (1.9)	1.18 (1.5)	10	22	29	30	17	16	8	10	17	4	14
12. Carrying a heavy object of 10 pounds? (*disability-heavy object*)	***47.5 (0.5)***	***1.51 (3.7)***	***1.68 (4.4)***	***46.0 (0.5)***	***1.55 (4.0)***	***1.63 (4.3)***	32	15	33	26	8	11	12	7	20	3	7
13. Removing something from your back pocket? (*disability-back pocket*)	53.3 (0.6)	0.97 (-0.2)	0.79 (-1.3)	52.6 (0.6)	0.99 (0.0)	0.85 (-0.9)	72	26	24	12	15	8	7	3	7	4	0

Italicized text indicates item misfit, defined as Infit or Outfit statistic < 0.6 or > 1.4, z-score > 2.0

*Disability and Pain subscales analyzed separately

Appendix 10. NDII Summary Table

NDII Item (*item stem*)	**Full Scale**			**Response Category Frequency ***				
	Measure (Model S.E.)	**Item Infit Statistic (z-score)**	**Item Outfit Statistic (z-score)**	**1**	**2**	**3**	**4**	**5**
1. Are you bothered by neck or shoulder pain or discomfort? (*pain/discomfort*)	59.3 (1.1)	0.92 (-0.8)	0.94 (-0.6)	25	43	47	57	8
2. Are you bothered by neck or shoulder stiffness? (*stiffness*)	62.1 (1.1)	0.97 (-0.3)	0.97 (-0.3)	26	38	59	52	4
3. Are you bothered by difficulty with self-care activities because of your neck or shoulder? (*self-care*)	40.0 (1.1)	1.08 (0.8)	1.01 (0.1)	4	21	39	66	49
4. Have you been limited in your ability to lift light objects because of your shoulder or neck? (*lift-light objects*)	38.8 (1.1)	1.13 (1.1)	0.99 (0.0)	5	22	30	59	64
5. Have you been limited in your ability to lift heavy objects because of your shoulder or neck? (*lift-heavy objects*)	58.5 (1.0)	1.20 (1.9)	1.16 (1.5)	29	40	55	39	15
6. Have you been limited in your ability to reach above for objects because of your shoulder or neck? (*reach above*)	59.2 (1.0)	1.38 (3.3)	1.35 (2.9)	33	40	48	45	14
7. Are you bothered by your overall activity level because of your shoulder or neck? (*overall activity*)	50.8 (1.1)	0.73 (-2.9)	0.73 (-2.8)	12	35	49	63	21
8. Has the treatment of your neck affected your participation in social activities? (*participation-social*)	38.1 (1.0)	0.91 (-0.7)	0.78 (-1.2)	7	19	25	47	80
9. Have you been limited in your ability to do leisure or recreational activities because of your neck or shoulder? (*participation-leisure/recreation)*	46.1 (1.0)	0.84 (-1.5)	0.77 (-2.2)	13	26	37	60	43
10. Have you been limited in your ability to do work (including work at home) because of your neck or shoulder? (*participation-work*)	47.1 (1.0)	0.80 (-1.9)	0.74 (-2.3)	17	26	37	54	45

Italicized text indicates item misfit, defined as Infit or Outfit statistic < 0.6 or > 1.4, z-score > 2.0

*Recall response categories were flipped for analysis. Here 1 =Not at all, 2= a little bit, 3 = a moderate amount, 4= quite a bit, 5 = a lot