Reliability and Responsiveness of the Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)

by

Nadia J Luca, MD, FRCPC, FAAP

A thesis submitted in conformity with the requirements
for the degree of Masters of Science Degree

Institute of Health Policy Management and Evaluation
University of Toronto

Reliability and Responsiveness of the Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)

Nadia J Luca

Master of Science (Clinical Epidemiology and Health Care Research)

Institute of Health Policy, Management and Evaluation
University of Toronto

2013

Abstract

Aims: To determine the test-retest reliability and responsiveness of a new computerized 20-item pain measure, SUPER-KIDZ, in children with juvenile idiopathic arthritis (JIA).

Methods: A single centre prospective cohort study of JIA patients aged 8-18 years was performed. For each SUPER-KIDZ item, test-retest reliability analysis was done in patients expected to have stable pain, and responsiveness was evaluated after intra-articular steroid injection(s).

Results: Fifty-one subjects were included. Good internal consistency (α=0.73-0.92) was demonstrated for the 3 SUPER-KIDZ domains. Acceptable test-retest reliability (intraclass correlation coefficient or kappa $\geq$0.80) was found for 15 SUPER-KIDZ items. At 2 weeks post-injection, 16 items were responsive to change in pain (standardized response mean=0.66-0.82, significant Wilcoxon signed rank and linear mixed model).

Conclusions: The majority of the SUPER-KIDZ items have acceptable test-retest reliability and responsiveness properties. If validity is demonstrated, this measure could be implemented as a standardized comprehensive pain tool for JIA patients.

Acknowledgments

I would like to thank my supervisor Dr. Ahmed Bayoumi and the members of my thesis advisory committee Drs. Jennifer Stinson, Dorcas Beaton, Brian Feldman, and Susanne Benseler, for their invaluable guidance, mentorship, and support. I am grateful to have had the opportunity to work with experts in the fields of clinical epidemiology, pediatric rheumatology, and pain research. I also appreciate the advice from Dr. Annie Dupuis for the statistical analysis. Thank you to my reviewers, Dr. Martin Offringa and Dr. Adam Huber, for taking the time and interest to critically evaluate this thesis.

The division of Rheumatology has been so supportive of this study, and special thanks go to Elizabeth Morais-Rego who helped us to identify patients. In addition, Kristi Whitney championed the study in the clinic, and went above and beyond her role in helping us to identify and recruit many patients. I would also like to thank the patients and parents who gave their time and energy to participate in the study.

I would like to thank Margaret van Wyk for working tirelessly and enthusiastically as the clinical research assistant for the study. Only with her constant attention to detail and hard work were we able to collect excellent quality data.

To my parents and sisters, thank you for your ongoing support and encouragement throughout all of my studies. And to my future husband and best friend, Dev, thank you for your patience, for cheering me on, and for always believing in me.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AIC: Akaike's Information Criterion

ANOVA: Analysis of variance

AUC: Area under the curve

CARRA: Childhood Arthritis and Rheumatology Research Alliance

CBT: Cognitive behavioural therapy

CHAQ: Childhood Health Assessment Questionnaire

CI: Confidence Interval

CRA: Clinical research assistant

DMARD: Disease-modifying anti-rheumatic drug

ES: Effect size

GRCP: Global rating of change in pain

GRCS: Global rating of change scales

HRQL: Health-related quality of life

ICC: Intra-class correlation coefficient

ILAR: International League of Associations for Rheumatology

ITC: Item-total correlation

IQR: Interquartile range

JI: Joint injection

JIA: Juvenile idiopathic arthritis

MCID: Minimal clinically important difference

MDC: Minimal detectable change

NSAID: Non-steroidal anti-inflammatory drug

PedIMMPACT: Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials

PedsQL: Pediatric quality of life inventory

PGA: Physician global assessment

PPQ: Pediatric Pain Questionnaire

NRS: Numerical rating scale

NSAID: Non-steroidal anti-inflammatory drug

RF: Rheumatoid factor

ROC: Receiver operating characteristic

RR: Guyatt's responsiveness ratio

SEM: Standard error of measurement

SD: Standard deviation

SE_{diff}: Standard error of difference scores

SRM: Standardized response mean

SUPER-KIDZ: Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth

VAS: visual analogue scale

1 INTRODUCTION

1.1 Juvenile Idiopathic Arthritis

1.1.1 Epidemiology and Classification

Juvenile idiopathic arthritis (JIA) is the most common chronic rheumatic disease in childhood, with an annual incidence of approximately 7 per 100,000 children in Canada [1]. It is diagnosed in children younger than 16 years of age, with arthritis in one or more joints for at least 6 weeks, and other causes excluded [2]. Arthritis in JIA is characterized by stiffness, pain and swelling of affected joints. The disease course of JIA may involve flares of increased disease activity or chronic persistent joint inflammation, often into adulthood [3, 4]. Various complications may arise secondary to ongoing disease activity or treatment, including joint damage and deformity, growth abnormalities, and osteoporosis with fragility fractures [5-7].

JIA represents a heterogeneous group of diseases. The current International League of Associations for Rheumatology (ILAR) classification of JIA consists of seven subtypes, which are described in Table 1 [2]. The classification system is based on the predominant clinical and laboratory features within the first 6 months of disease. It was developed to distinguish relatively homogeneous, mutually exclusive categories of idiopathic childhood arthritis to aid in the conduct of research studies. The grouping may also predict, to some extent, homogeneity of response to therapies.

1.1.2 Management of JIA

The approach to the management of JIA is multi-modal, involving both pharmacologic and non-pharmacologic therapies. The goal is to intervene as early as possible to reduce joint inflammation, and ultimately prevent joint damage and preserve daily functioning. Although there is no “cure” for JIA, disease remission is a realistic outcome.

Drug therapy typically starts with non-steroidal anti-inflammatory drugs (NSAIDs), with the addition of a disease-modifying anti-rheumatic drug (DMARD) such as methotrexate often necessary for disease control. Corticosteroids, usually intra-articular and occasionally systemic, may be used. Intra-articular injection of triamcinolone-hexacetonide has been shown to have positive and long lasting effects in JIA [8]. If patients fail to respond to these therapies,

biologic agents such as anti-tumor necrosis factor alpha agents (e.g. etanercept) are introduced. These agents target key cytokines implicated in the pathogenesis of JIA. Response to therapy has been defined as percent improvement in at least 3 of 6 core response variables including: physician global assessment of disease activity (10-cm visual analogue scale [VAS]), parent or patient global assessment of well-being (10-cm VAS), number of active joints, number of joints with reduced range of motion, functional ability, and erythrocyte sedimentation rate [9].

Non-pharmacologic therapy includes physiotherapy with range of motion exercises, stretching, and strengthening. Aerobic exercise is felt to be beneficial and, although a therapeutic effect has not been demonstrated, it is certainly felt to be safe for children with JIA [10, 11]. In addition, the use of splinting and orthotics can correct deformities, malalignment, and decrease pain [12]. Nutrition is another important component of overall management, particularly ensuring optimal calcium and vitamin D intake to reduce the risk of osteoporosis [13].

Table 1: ILAR Classification of Juvenile Idiopathic Arthritis [2, 14]

Subtype	**ILAR Criteria**	**Prevalence**
Systemic	Arthritis in ≥1 joint, quotidian fever x ≥2 weeks and ≥1 of the following: i) evanescent (nonfixed) erythematous rash ii) enlarged liver or spleen iii) enlarged lymph nodes iv) serositis	4-17%
Olgoarticular Persistent Extended	Arthritis in 1-4 joints during first 6 months of disease ≤4 joints throughout disease course >4 joints after first 6 months of disease	27-56%
Polyarticular (RF-)	Arthritis in ≥5 joints during first 6 months of disease And negative test for RF	11-28%
Polyarticular (RF+)	Arthritis in ≥5 joints during first 6 months of disease And ≥2 positive tests for RF at least 3 months apart	2-7%
Psoriatic	Arthritis and psoriasis OR arthritis and ≥2 of: i) dactylitis ii) nail pitting or onycholysis iii) psoriasis in 1st degree relative	2-11%
Enthesitis-related	Arthritis and enthesitis OR arthritis or enthesitis with ≥2 of: i) sacroiliac joint tenderness and/or inflammatory lumbosacral pain ii) HLA-B27 antigen iii) onset of arthritis in a male >6 years iv) acute (symptomatic) anterior uveitis v) history of ankylosing spondylitis, enthesitis-related arthritis, sacroiliitis with inflammatory bowel disease, Reiter's syndrome, or acute anterior uveitis in a 1st degree relative	3-11%
Undifferentiated	Arthritis that fulfills criteria in no category or in 2 or more of the above categories	11-21%

ILAR= International league of associations for rheumatology; RF = rheumatoid factor

1.2 Pediatric Pain

Pain is defined as an "unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" [15]. Pain is always subjective. Each individual learns the application of the word through experiences related to injury in early life [16]. Acute pain can arise from medical procedures, injury, or acute exacerbations of disease pain. Chronic or recurrent pain is defined by the American Pain Society as any prolonged pain that lasts longer than the expected healing time (arbitrarily defined as greater than three to six months), or any recurrent pain that occurs at least three times throughout a period of three months [16]. By its nature, chronic or recurrent pain is often difficult to manage [17]. Acute and chronic pain may occur concurrently. For example, children with arthritis might experience an acute flare in their pain while living with persistent chronic arthritic pain.

The underlying construct of pain includes a vast variety of qualities and experiences, not only pain intensity. Melzack and colleagues eloquently emphasize the complexity of pain and show that it consists of 3 dimensions: a) sensory-discriminative, b) affective-motivational, and c) cognitive-evaluative [18, 19]. The sensory-discriminative dimension refers to the intensity, quality and location of the pain. The affective-motivational aspect captures how one feels when in pain, the emotional arousal related to pain and its aversive effects. The cognitive-evaluative dimension reflects cognitive pain behaviours, suffering, and the individuals' perceptions of the influence of pain on aspects of their health-related quality of life (HRQL) (i.e., pain's interference with aspects of physical, psychological, and social functioning). While pain intensity alone can be a sensitive measure when evaluating pain treatment outcomes [20], assessing the other dimensions yields a more complete understanding of the pain experience and how to better address its treatment and control.

Given the multi-dimensional nature of pain, the Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (PedIMMPACT) group has described the core domains that should be measured as outcomes in clinical trials of pediatric chronic or recurrent pain [21]. They propose that the following 8 domains be considered: pain intensity, physical functioning, emotional functioning, role functioning, sleep, economic factors, symptoms and adverse events, and global satisfaction with pain treatment [21]. A key point is that pain should

be assessed comprehensively, including some determination of the extent of pain interference, and, as such, a pain measure should capture this multi-dimensionality (section 1.3.5).

1.3 Pain in JIA

1.3.1 Prevalence of Pain in JIA

Pain is the most common and distressing symptom in JIA [22, 23]. There has been significant debate in the literature about pain in JIA owing to early reports that children with JIA experience less pain than adults with rheumatoid arthritis [24, 25]. However this is likely due to the fact that young patients express pain differently compared with adults, and that they were being assessed with developmentally-inappropriate pain measures [26]. Several studies have clearly shown that JIA patients do in fact experience significant pain. For example, a Norwegian study reported that 82% of JIA patients report daily pain lasting a mean of 4.3 hours [27]. In an American study of children with polyarticular JIA, 76% of children reported pain on more than 60% of days despite being treated with disease-modifying therapy [28]. Seventy-seven percent of North American pediatric rheumatologists acknowledge that children with arthritis continue to have clinically significant pain despite adequate doses of NSAIDs and DMARDs [29].

1.3.2 Etiology and Predictors of Pain in JIA

The etiology of pain in patients with JIA is multifactorial, and is not completely understood. Contributing factors include the disease process itself, which may lead to peripheral and central sensitization, as well as cognitive-behavioural, emotional, and environmental influences.

Joint inflammation results in disruption of cellular membranes, and creation of arachidonic acid and prostaglandins (eg. PGE1, PGI2), which sensitize pain nociceptors in the joint capsule and periosteum [30]. The pain signal is initiated by histamine and bradykinin, both of which are also released during inflammation. Small unmeylinated sensory nerves carry the pain signal to the dorsal ganglia and higher order neurons of the central nervous system. Peripheral sensitization may occur due to local factors at the site of the inflamed joint [31]. Synovial fluid, tissue edema and increased blood flow lead to increased intra-articular pressure, which further triggers the nociceptor fibres. In addition, previously silent nociceptive fibres become involved

due to the inflammation. Nerves in the synovium also contain substance P, which not only contributes to pain, but may also contribute to maintaining the inflammatory process.

Interestingly, pain reported by children with JIA does not appear to be fully explained by disease activity and inflammation [32-34]. Studies have suggested that JIA patients may develop a generalized reduced pain threshold. Hogeweg et al showed that patients with active arthritis have a significantly lower mean pain threshold than JIA patients in remission, but that both groups have lower pain thresholds compared with healthy controls [35]. Pain thresholds were found to negatively correlate with pain VAS and the functional component of the child health assessment questionnaire (CHAQ). Thus, it appears that 'central sensitization' may play a role in the pain experience in JIA. That is, repetitive nociceptive input from joint inflammation produces alterations in the response properties of second-order neurons in the dorsal horn, which results in an increased responsiveness within the spinal cord [36].

There is also evidence for the contribution of cognitive-behavioural factors to pain perception in patients with JIA. Schanberg et al have demonstrated that increased use of pain coping strategies (e.g. distraction) and decreased catastrophizing (i.e. negative thinking about pain) lead to significantly lower pain ratings [23]. Cognitive health beliefs such as increased disability and harm, and decreased control also appear to be important predictors of increased pain [37, 38]. In addition, data show that daily fluctuations in mood and stress level significantly impact reported pain, stiffness and fatigue [39]. Similarly, impaired regulation of negative emotions in children with JIA has been shown to predict higher reported pain intensity and functional limitation [40]. Environmental influences such as increased parental pain experiences tend to correlate with higher reported pain in their children [41]. Also important is parental response to their child's pain, with studies showing that responses promoting self-management and coping are more beneficial than those reinforcing illness behaviours in children [42]. These findings support the idea that disease activity accounts for only part of the pain experienced by children with JIA, and that cognitive-behavioural, emotional, and environmental factors should be considered when measuring and managing pain in this population.

1.3.3 Consequences of Pain in JIA

Pain is associated with decreased HRQL in children with JIA [43-47]. Interestingly, HRQL in children with JIA appears to be primarily explained by pain intensity, functional disability, missing school and burden of taking medications, rather than with disease activity level [47]. One study showed that pain intensity of >3.4/10 is one of the strongest predictors of poor physical well-being (odds ratio [OR] = 2.5 [1.8-3.5]), and is the single strongest predictor of poor psychosocial well-being (OR=4.7 [2.0-7.6]) [44]. Greater pain intensity is associated with greater functional disability and activity limitation (correlation=0.67) [40, 48], which in turn also impacts HRQL [44, 46].

Pain and discomfort can impede self-care activities, impair leisure and physical activities, and disrupt school attendance and sports participation [28]. Increased pain is also correlated with higher scores on depression scales [49], as well as poor sleep and fatigue in patients with JIA [50, 51], and these relationships may be bidirectional. Even a small reduction in pain (i.e. 0.82 cm on 10 cm VAS) is associated with improved quality of life in JIA patients [52]. A recurrent theme in qualitative studies is that unrelenting and unpredictable pain leads to youth with JIA feeling different, fearful, frustrated and powerless [53]. In addition, many youth report frustration and hurt due to lack of empathy and understanding from others who cannot see their "invisible" arthritis pain and symptoms [53].

1.3.4 Management of Pain in JIA

There are no evidence-based guidelines specific to the management of acute or persistent pain in JIA. Reviews on this topic suggest a multi-disciplinary approach starting with early and aggressive control of underlying joint inflammation. Additional pharmacologic therapy may include the use of analgesics such as NSAIDs and acetaminophen, and consideration of opioids in particularly refractory cases [17, 54-56]. Warm packs, splinting and orthotics can be helpful in reducing pain and improving mobility. Sleep hygiene is also important given the relationship of sleep disturbance and increased pain in children with JIA [50, 51]. If accessible, cognitive behavioural therapy (CBT) or other psychological therapies may be very effective in reducing chronic pain in patients with JIA [57].

1.3.5 Measuring pain in JIA

As discussed above, the construct of pediatric pain is multidimensional and complex, and these characteristics should be captured by a potential pain measure. The measure should consist of multiple domains that address the type and intensity of pain, factors contributing to pain, as well as its impact on function and quality of life.

In the context of JIA, the purpose of measuring pain is to detect clinically significant changes in pain over time and in response to various interventions to enable optimal pain management. Thus, according to the methodological framework proposed by Kirshner and Guyatt [58], a pain assessment tool is 'evaluative' in nature. In this case it should contain items that are likely to change in response to an intervention, and when clinically important improvement or deterioration occurs. All clinically important effects related to the outcome should be included, and items with multiple response options (i.e., 7-10 point scales or VAS) are best to enable detection of change. If possible, items that are unresponsive to change should be avoided.

A pain measure should be consistent with the age and cognitive development of the child. The term 'pain' is an abstract concept, which may be meaningless to children in the pre-abstract stage of cognitive development [26]; however, when other words (e.g. sore, aching, stinging) are used, these sensations are endorsed [59]. At younger than 4 years old, children are not able to report about pain intensity or negative emotions from pain [60]. Children aged 4 and older are able to self-report, but those aged 4-7 years can typically report only on pain intensity and location [20, 61]. Thus, parent and or health care provider proxy-reports are often necessary for this age group. Older children (≥8 years) are able to use VAS scales and to report on the impact of pain in their daily life [20].

1.3.6 Existing Pain Measures for JIA

This section briefly reviews the pain measures that exist for use in JIA population (Table 2).

1.3.6.1 Pediatric Pain Questionnaire

James W. Varni and Karen L. Thompson developed the Varni/Thompson Pediatric Pain Questionnaire (PPQ) in 1985 for the assessment of musculoskeletal pain in children aged 4-16 years. Interestingly, it was never officially published [62]. The measure consists of 6 items, including pain intensity (VAS), a body outline for location, and asks the child to choose words

that best describe their pain and how they feel when in pain. The parent version includes a question about pain interference. Eleven years after its development, the researchers carried out validation studies for test-retest reliability (over 6 months) and construct validity in children with rheumatic diseases aged 8-16 years [63]. However, only the pain intensity ratings were tested, and no a priori hypotheses were made. Feasibility and responsiveness have not been evaluated. The PPQ does not appear to be routinely used in clinical practice [17].

1.3.6.2 e-Ouch© Multidimensional Pain Diary

The e-Ouch© multidimensional pain diary is an electronic pain diary accessible via a portable handheld device and is designed to obtain pain ratings from adolescents with JIA 3-times per day [64]. The e-Ouch© not only collects data related to pain intensity, but also pain unpleasantness, pain interference (i.e., impact of pain on mood, sleep, walking, relationships and enjoyment of life) and perceived control over pain. By virtue of its electronic data-capture method allowing for immediate ratings of current pain status, it is not subject to recall bias. The e-Ouch© diary has been shown to be a valid and responsive assessment modality [65], although reliability has not been assessed. However, analysis of the data from the eOuch© diary requires complex statistical analyses [66]. To date this tool has only been used in clinical research and has not be employed to track musculoskeletal pain in routine clinical practice.

1.3.6.3 Pediatric Quality of Life Inventory and Childhood Health Assessment Questionnaire

The Pediatric Quality of Life Inventory (PedsQL-Rheumatology module) is a validated measure for HRQL in the pediatric rheumatology population [67]. It contains 4 items evaluating ‘pain and hurt’, but it is meant to be used as a broader assessment of all components of HRQL, not only pain.

The Childhood Health Assessment Questionnaire (CHAQ) primarily measures the ability of a child to function in 8 areas of daily life (disability index) with one question about pain intensity on a 100 mm VAS (discomfort index) [68]. Reliability and construct validity have been established for the disability index but not the discomfort index.

While they provide valuable information about HRQL and functional ability, the PedsQL and CHAQ on their own are inadequate for comprehensive pain assessment as they assess only pain intensity, and not the affective or evaluative aspects of the pain experience.

1.3.6.4 Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)

The above pain measures for children with JIA are either not comprehensive (PPQ, PedQL, CHAQ) or not developmentally appropriate for the full pediatric age-range (PPQ, eOuch©). In response to this gap in clinical care, the Childhood Arthritis and Rheumatology Research Alliance (CARRA) (a group of 280 pediatric rheumatologists across North America committed to research for the prevention and therapy of childhood rheumatic diseases) supported its Pain Subcommittee to develop a new comprehensive online pain measure for children aged 4-18 years with rheumatic disease, called the Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth **(**SUPER-KIDZ). This multi-dimensional tool consists of 3 developmentally appropriate versions. The development and feasibility testing of the SUPER-KIDZ tool is described in detail in section 1.5. The focus of the current study is to determine the reliability and responsiveness of the SUPER-KIDZ pain measure in the JIA population. As such, measurement theory is first presented in section 1.4.

Table 2: Summary table of pediatric pain measures for children and youth with JIA ([66])

Measure	Content	Respondent burden	Adminstra-tive burden	Reliability	Validity	Responsive-ness	Strengths	Cautions
PPQ	Pain intensity (2 items, VAS), body outline, words describing pain (3 items). Parent version has interference item	10-15 min	5 min to hand score	'Moderate stability' with simple correlation (VAS, 8-16 yrs only)	Concurrent validity established (VAS, 8-16 yrs only)	Not assessed	Body location includes pain intensity scoring	Not development-ally appropriate for <8 years. No functional questions in child version.
eOuch© electro-nic pain diary	Sensory, interference, unpleasantness, fatigue, stiffness, control over pain (VAS scales)	3 daily pain ratings, total 9 min	Computer algorithm for scoring	Not assessed	Construct validity established	Detected change in pain in adolescents undergoing joint injection	Comprehensive, no recall bias, feasible in home setting	Challenging data analyses. Only for adolescents.
PedsQL	Pain and hurt scale (4 items, ordinal scale)	Whole PedsQL takes 10 min	Scored by hand	Cronbach's alpha	Construct validity established	Detected change in patients undergoing treatment	Developmentally appropriate for ages 4-18	Primarily intended for HRQL assessment
CHAQ	Discomfort index: pain intensity VAS	Whole CHAQ takes 5-10 min	Scored by hand	Only for disability index	Only for disability index	Not assessed	Parent-proxy version available	Primarily intended for functional assessment
SUPER-KIDZ	Sensory (NRS), fatigue, cognitive-evaluative, affective (ordinal scales)	3-4 min, complete online	Paper print-out; scoring algorithm TBD	Test-retest reliability under study	Construct validity under study	Responsive-ness after joint injection under study	Comprehensive; developmentally appropriate for ages 4-18; feasible in home setting	Still need to determine scoring

PPQ=pediatric pain questionnaire; PedsQL=pediatric quality of life inventory-rheumatology module; CHAQ=child health assessment questionnaire; SUPER-KIDZ=standardized universal pain evaluations; VAS=visual analog scale; NRS=numerical rating scale

1.4 Concepts of Measurement

Measurement of a health state is the cornerstone of clinical practice and medical research. As stated in 1883 by Lord Kelvin, "when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot, your knowledge is of a meagre and unsatisfactory kind" [69]. Prior to developing or choosing a clinical measurement instrument, one must clearly specify the underlying construct to be measured, the purpose of measuring it, and the population in which you want to measure it (as described in Section 1.3.5). Once it is developed, the tool must be validated.

1.4.1 Validation of a health status measure

Prior to its implementation into clinical practice, a health status measure should be pre-tested in a small group of users to ensure comprehensibility and relevance [70]. Feasibility should be assessed, with consideration given to ease of use, and minimal burden on the patient and health care provider. Without these characteristics, the tool is unlikely to be accepted by medical practitioners or adopted into clinical care.

Measurement instruments must also be assessed with regards to several measurement properties including reliability, validity, responsiveness and interpretability (Table 3) [71]. Briefly, reliability may refer to properties such as internal consistency, reliability (inter-rater or test-retest), or measurement error. The goal is to determine the amount of variance of a measurement that is due to systematic and random error, as opposed to "true" variance between patients. Face and content validity are an assessment of the reasonableness of the items in the measure, and whether they cover all aspects of the underlying construct. Construct validity evaluates hypothesized relationships of the measure under study and other measures, and criterion validity is assessed when there is a gold standard for comparison. Responsiveness refers to whether an instrument is able to detect change when it has occurred, while interpretability is the ability to assign significance to the measure's scores or change in scores.

Reliability and responsiveness are discussed in more detail in sections 1.4.3 and 1.4.4 below since they are the focus of the current study.

Table 3: Definitions of domains of measurement properties. Adapted from COSMIN taxonomy [71].

Domain	Measurement property	Definition
Reliability	Internal consistency	The degree of interrelatedness among items
	Reliability	The proportion of the total variance in measurements which is due to "true" differences among patients
	Measurement error	The systematic and random error of a patient's score that is not attributed to true changes in the construct being measured
Validity	Content validity	The degree to which the content of an instrument is an adequate reflection of the construct to be measured
	Face validity	The degree to which the items appear as an adequate reflection of the construct to be measured
	Construct validity	The degree to which the scores of an instrument are consistent with hypotheses (e.g. relationships to scores of other instruments) based on the assumption that the instrument validly measures the construct to be measured
	Criterion validity	The degree to which the scores of an instrument are an adequate reflection of a "gold standard"
Responsiveness		The ability of an instrument to detect change over time in the construct to be measured
Interpretability		The degree to which one can assign qualitative meaning to an instrument's quantitative scores or change in scores

1.4.2 Reliability and Related Concepts

1.4.2.1 Internal consistency

Internal consistency is defined as the degree of inter-relatedness among items in a unidimensional scale or subscale of an instrument [71]. It is a measure of the extent to which items assess the same construct. The item-total correlation (ITC) and Cronbach's alpha (α) are commonly used parameters for assessing the internal consistency of a scale.

ITC is the correlation of each item with the sum of the remaining items [72]. It gives an indication of whether an item can discriminate between patients with higher or lower levels of

the construct being measured. If an item has an ITC of less than 0.3, it is not able to distinguish between mildly and highly affected patients [73].

Cronbach's α measures the correlation between the items on the (sub)scale, to see how much they are associated [72]. Calculating Cronbach's α as items are sequentially eliminated from the scale can help to determine which items might *not* be well correlated with the other items in the scale (i.e. the α would increase). An accepted guideline for the value of Cronbach's α is between 0.70 and 0.95; a value higher than 0.95 may indicate redundancy of items [74]. It is important to note that the value of Cronbach's α is highly dependent on the number of items in the scale. With a large number of items, it may have a high value, despite rather low inter-item correlations [74].

Cronbach's α is based on a Pearson covariance matrix, and if the data are not continuous the matrix may be distorted. If assumptions are violated, the reliability estimate can be substantively deflated. In the case of non-continuous data, it is preferable to use a polychoric correlation matrix to calculate an ordinal reliability coefficient alpha (α) which gives a more accurate estimate of α for ordinal data [75].

1.4.2.2 Reliability

There are several sources of variance in a measurement including biological variability, instrumentation, error by the subject, and error by the tester. Reliability is defined as the proportion of the total variance in measurements that is due to the "true" differences between subjects [74], and can generally be calculated as follows [76]:

$$\text{Reliability} = \frac{\text{between-subjects variance}}{(\text{between-subjects variance} + \text{error variance})} \quad \text{(Equation 1a)}$$

or $$\text{Reliability} = \sigma_S^2 / (\sigma_S^2 + \sigma_E^2) \quad \text{(Equation 1b)}$$

Where σ_S^2 = between-subjects variance, σ_E^2 = error variance

Thus, if measurement error is small in comparison with variability between subjects, the reliability parameter approaches 1. For continuous variables, reliability can be quantified by the intra-class correlation coefficient (ICC), wherein the error variance can be differentiated into 'facets of interest', such as variance between raters (as in inter-rater reliability) or trials (as in test-retest reliability) [72]:

$$ICC = \sigma_S^2 / (\sigma_S^2 + \sigma_T^2 + \sigma_{re}^2) \quad \text{(Equation 2)}$$

Where σ_S^2 = between-subject variance, σ_T^2 = variance of facet of interest (e.g. rater, trial) and σ_{re}^2 = residual error variance

The necessary variance estimates for the calculation of the ICC are derived from a repeated-measures analysis of variance (ANOVA). Shrout and Fleiss describe 3 models of reliability parameters according to how the raters (or trials) are chosen and assigned to the subjects (total of 6 ICCs) (Table 4) [77]. In order to choose the correct ICC, the first decision is whether a one- or two- way ANOVA is appropriate for the study design. If a different rater (trial) judges each subject (i.e. no interaction between the subject and the raters [trials]), then a one-way ANOVA is appropriate, and the overall variance is simply partitioned into between-subject and within-subject variance (Model 1). The ICC calculation follows Equation 1b above [76]. However, if the same rater (trial) judges each subject, they are not independent (Models 2 and 3). In this case, a two-way ANOVA is required to account for the interaction between subjects and raters (trials) by partitioning the within-subject variance into a between-rater (trial) variance and a residual error variance. The ICC calculation then follows Equation 2 above. A second consideration is whether the factor being studied (raters or trials) is considered a random-effect or a fixed-effect [77]. In the former case, the factor is a random sample of the possible levels of this factor (Model 2), while in the latter only a single level or fixed set of levels is of interest (Model 3). Mathematically, Model 2 assumes that both a) the differences between the raters' mean score and population mean and b) the interaction components for each subject and rater, are mutually independent and normally distributed, whereas Model 3 constrains that the sums of these variables be zero. Lastly, for each of the three models, there is a choice depending on whether one is interested in the reliability of a single rating or of a mean of several ratings. The latter simplifies the ICC calculation and is always greater in magnitude than the ICC for single scores [76, 77].

For the purposes of our test-retest reliability study, the factor of interest is the effect of the trial (administration of the SUPER-KIDZ measure), which is done the same way for each subject. That is, there may be a systematic effect or interaction between the manner in which the questionnaire is administered and the subjects' responses, and they are not independent (thus Model 1 is not appropriate). In addition, this set of trial circumstances is not the only type of

interest, as we wish to implement the questionnaire into all clinical settings. Thus, a fixed-effect model (Model 3) is inappropriate. Lastly, we are interested in the reliability of single scores, not means. According to the above discussion, the ICC of choice is a two-way random-effects model or Shrout and Fleiss ICC (2,1) [76]. A two-way ANOVA partitions the within-subject variance into variance due to the trial and the unexplained error variance. The random-effects layout means that the interaction component of the variance contributes additively to each expected mean squares [77]. The mean squares from the ANOVA are used to calculate the ICC parameter as follows:

$$ICC\ (2,1) = (BMS - EMS) / (BMS + (k-1)*EMS + k(RMS-EMS)/n) \qquad \text{(Equation 3)}$$

Where BMS = between-subjects mean square, EMS = error mean square, RMS = between-trials mean square, k = number of trials, n = number of subjects tested

Table 4: Summary of intraclass correlation coefficient (ICC) models (adapted from Weir, 2005 [76])

Shrout & Fleiss nomenclature		**Type of ANOVA**	**Single or mean scores**
Model 1	ICC (1,1)	1-way random	Single
	ICC (1,k)	1-way random	Mean
Model 2	ICC (2,1)	2-way random	Single
	ICC (2,k)	2-way random	Mean
Model 3	ICC (3,1)	2-way fixed	Single
	ICC (3,k)	2-way fixed	Mean

Where k=number of ratings used to form the mean

Interpretation of the ICC is controversial and universal standards are difficult to define. Generally an ICC of 0.70 to 0.75 is acceptable for interpretation of scores at a group-level [73, 78]. Higher values, ideally exceeding 0.9, are typically required for individual decisions [72]. However, in the case of pain, which is a dynamic construct that can fluctuate over hours, an ICC of 0.8 may be more realistic [61, 79]. It should also be noted that, given the way it is calculated, the magnitude of the ICC is dependent on the degree of variability in the data. Thus, a lack of variability in subjects' scores will result in a smaller ICC [72].

For categorical variables or ordinal data, an ICC cannot be calculated, and reliability is assessed by measuring agreement [72]. Percent agreement is calculated as the number of exact agreements divided by the number of possible agreements. However, it does not take into account that some portion of the agreements may occur by chance. A more meaningful

parameter is Cohen's kappa, which assesses agreement between scores while adjusting for agreement expected by chance. Thus the magnitude of kappa is generally smaller than percent agreement. Cohen's kappa is calculated by subtracting the expected agreement (P_e) from the observed agreement (P_o) and then dividing by the amount of agreement that can be maximally reached beyond chance [74]:

$$\kappa = (P_o - P_e) / (1 - P_e) \quad \text{(Equation 4)}$$

When there are more than 2 ordered categories a weighted Cohen's kappa is appropriate to more heavily penalize misclassifications between more distant categories. If a cross-table is constructed with *i* scores in rows and *j* scores in columns, the weighted kappa is calculated as [74]:

$$\kappa = 1 - ([\Sigma w_{ij} \times P_{oij}] / [\Sigma w_{ij} \times P_{eij}]) \quad \text{(Equation 5)}$$

where summation is taken over all cells (*i,j*), w_{ij} is the weight assigned to cell (*i,j*), and P_{oij} and P_{eij} are the observed and expected proportions of cell (*i,j*), respectively. The most common way to assign weights is to treat the scale as an ordinal continuum with equal intervals ('incremental weights') and apply weights in a linear or quadratic manner. Weights can be applied symmetrically or asymmetrically [72].

Kappa values range between -1 and 1. A kappa of 0 means there is no more agreement than can be expected by chance. Generally values above 0.4 are considered to be moderate [80]. Values of kappa depend on the heterogeneity of the sample, and lower kappa values may result in more homogeneous populations.

1.4.2.3 Error parameters

The standard error of measurement (SEM) is the standard deviation (SD) around a single measurement, and provides an absolute index of reliability at the individual subject level [72, 76]. It is related to reliability in that, with a more reliable measure, the distribution of errors is less variable. In contrast to the ICC, it has the benefit of not being affected by between-subjects variability, and carries the same units as the measurement of interest. The SEM can be estimated from the corresponding ANOVA as follows [76]:

$$SEM = SD \sqrt{(1-r_{xx})} \quad \text{(Equation 6)}$$

Where SD is the standard deviation of the observed test scores at baseline ($\sqrt{SS_{TOTAL}}/(n-1)$ and r_{xx} is the reliability coefficient (ICC) for that measurement.

It follows that the standard error of the difference scores (SE_{diff}) is equal to the SEM multiplied by square root 2 since it is derived from 2 test scores [81].

$$SE_{diff} = SD\ \sqrt{2}(1-r_{xx}) \qquad \text{(Equation 7)}$$

The SEM or SE_{diff} is used to define the minimal difference between individual measurements that can be considered "real" or exceeding the measurement error. If we assume that error follows a normal distribution, this value, the minimal detectable change (MDC) at a 95% confidence level, can be calculated using the formula [76]:

$$MDC_{95} = \sqrt{2} \times SEM \times 1.96 \qquad \text{(Equation 8a)}$$

$$MDC_{95} = SE_{diff} \times 1.96 \qquad \text{(Equation 8b)}$$

Thus, for all subjects whose difference in score on repeated testing is greater than or equal to the MDC_{95}, we can be 95% confident that they are not stable patients, given the measurement error of the test.

1.4.3 Responsiveness

Several definitions for responsiveness are reported in the literature, including: an ability to measure any change in state ("sensitivity to change") [82], an ability to measure a *clinically important* change (requires judgment of clinical importance) [83], or an ability to detect *real* changes in the construct being measured (requires a "gold standard") [84]. A generally accepted definition of responsiveness is "the ability of an instrument to accurately detect change when it has occurred" [85]. Some researchers view responsiveness as part of construct validity, i.e. testing the validity of the change score in a longitudinal design [83]. It is important to note that responsiveness is not a fixed property of an instrument, but rather it involves validating the application of the instrument in a specific test situation.

Evaluation of responsiveness requires that some form of change has occurred. In order to interpret responsiveness studies, the construct of change being quantified should be identified to allow responsiveness to be placed in the context of a specific type of change. It is important

that responsiveness be demonstrated in a context that is consistent with that in which the instrument will ultimately be used to measure change. The change construct has been defined by the 3-axis classification system outlined by Beaton and colleagues (Table 5) [84, 86]. Following this framework, one can categorize the type of change they wish to measure according to the 'who', 'which', and 'what' axes. From there, the appropriate study design and analysis can be established. The first 2 types of change in the "what" axis are related to the structure of the scale and boundaries of error. The third type of change has also been called "internal responsiveness" by Husted et al., and refers to an ability to measure observed change over time [87]. The latter 2 have been classified as "external responsiveness", and require a judgment and/or external measure to determine whether a patient has improved and whether this is an important improvement from either the patient's or the provider's perspective [87].

Table 5: Axes defining construct of change being measured (adapted from Beaton et al [84])

Axis	**Who** are the results presented for?	**Which** scores are being contrasted?	**What** type of change is being quantified?
Options	• Group-level • Individual-level	• Between persons at one point in time • Within person change over time • Between- and within-person changes (e.g. RCT)	• Minimum potentially detectable change by instrument • Minimum change detectable given measurement error • Observed change in a given population • Observed change in population deemed to have improved • Observed change in population deemed to have an important improvement

There are several statistical methods available to determine the magnitude of a change in health in subjects (Table 6). Most reflect a standardized ratio of observed change to variance. When compared to each other, different responsiveness indices give different results, and there is no consensus on the preferred index [82, 88, 89]. Guyatt's responsiveness ratio (RR) is defined as the minimal clinically important difference (MCID) divided by the variance in stable subjects [90]. However, it is limited in practicality because of the need to define the MCID. Effect size (ES) is a commonly used responsiveness statistic calculated by dividing the mean difference in

scores ($test_1$-$test_2$) by the SD of baseline scores [91], and the ES index is similar, dividing the mean difference by the SD of $test_2$ scores. These parameters do not incorporate information about the variance of the change scores. The standardized response mean (SRM) is calculated by dividing the mean difference in scores by the SD of the difference scores [92]. The ES and SRM can be interpreted as small (0.2-0.49), moderate (0.5-0.79), and large (≥0.8) [92]. However, it is important to keep in mind that the magnitude of the ES and SRM values is dependent on the type of change being examined, and hypotheses should be specified accordingly [93, 94]. The paired t test (or Wilcoxon signed rank test for non-parametric data) can also be used to determine whether the change in scores is statistically significant, but is influenced by the sample size.

If there are multiple time points, a repeated-measures ANOVA can be used to determine the main effect of time [95, 96], however this analysis assumes homogeneity of variances and does not provide information on the magnitude of change. A regression model is more flexible as it can be generalized to non-linear forms, control for potential confounders, be assessed for goodness-of-fit, and provide information on the magnitude of change in scores through the beta coefficients [87, 97]. A linear mixed model is a good choice for repeated measures because it accommodates heterogeneous variance structures, and is robust for missing data and varying numbers and timing of time points among subjects [98].

The above methods can be applied to all subjects expected to improve (internal responsiveness) or to those patients who are judged to have improved or not improved (external responsiveness), thereby incorporating information from an external clinical or health status measure [87]. In this case there should be a hypothesis about the expected change and the responsiveness statistic for those who have reported improvement versus those who have not improved. There are also several other methods to assess external responsiveness. One approach is to use an independent t test or ANOVA to compare the change in patients who undergo an important change with the change in patients who are not expected to change [94]. Another approach is to view a scale's responsiveness as its ability to discriminate between those who improve and those who do not, similar to a diagnostic test [99]. An external criterion such as a global rating of change scale (GRCS) (section 1.4.6) is used as the "gold standard" to categorize patients according to improvement. A receiver operating characteristic (ROC) curve is constructed by calculating the sensitivity and specificity of several potential change scores

for defining improvement compared with the gold standard. The area under the curve (AUC) represents the responsiveness, thus a larger AUC suggests a more responsive instrument for that situation.

Table 6: Description of responsiveness statistics used to measure change (from [83, 87, 94])

Type	Responsiveness statistic	Calculation	Comments
Internal (change expected due to treatment effect)[§]	Guyatt's Responsiveness Ratio (RR)	RR=MCID/SD of change scores in stable patients	Must be able to determine MCID
	Effect size (ES)	ES=average change/SD of initial scores	Does not incorporate variation in change scores
	Standardized response mean (SRM)	SRM=average change/SD of change scores	Must have a priori hypothesis
	Paired t test (or Wilcoxon signed rank test*)	t=average change/(SD of change scores/√n)	Similar to SRM but significance depends on sample size
	Repeated measures ANOVA	*F* statistic and p-value for main effect of time	No magnitude of change
	Regression model	p-value for association between measure and time	Beta coefficient=change in measure per increment of time
External (gold standard= change according to patient or provider)	Independent t test or ANOVA of change scores of 2 groups	t or *F* statistic and p-value	Differentiate among patients who are expected to change or not
	Receiver operating characteristic (ROC) curve area	Plot of sensitivity versus 1-specificity	Patients categorized into 2 groups according to external criterion of improvement
	Correlation coefficient[¥]	Pearson's r or Spearman's ρ*	Correlation between change scores and change in external measures of health status
	Regression model	p-value for association between measure and gold standard	Beta coefficient=change in measure per unit change in gold standard

[§]Can also measure external responsiveness if applied to patients classified as improved or not improved; [¥]often referred to as "longitudinal construct validity"; *if non-normally distributed data. SD=standard deviation; MCID=minimally clinically important difference; ANOVA= analysis of variance

Correlation analyses are also used to assess responsiveness, consistent with the idea that responsiveness be viewed as a component of construct validity (i.e. longitudinal construct validity). In this case, a Spearman or Pearson correlation is calculated between the change

scores of the tool being studied and other scores known to represent a change in the clinical state. Thus correlations are inherently measures of external responsiveness. As with construct validity, it is important to indicate a priori hypotheses for the expected correlations [78, 99].

As noted above, regression models provide a lot of information regarding the responsiveness of a measure, not only whether there is a significant difference in scores over time, but also an estimate of the magnitude of change in score per unit time. These models can also be used to assess the external responsiveness of a measure by testing whether change in the gold standard measure predicts change in the measure under study (or vice versa) [87, 100, 101]. In this case, the beta coefficient represents the magnitude of change in the measure under study corresponding to a one-unit change in the gold standard.

Often researchers will test a few constructs of change and use several statistical methods to demonstrate responsiveness. For example, a study by Krahn et al evaluated the responsiveness of utility measures in prostate cancer patients [100]. They determined the ES and SRM (internal responsiveness) in the population expected to improve after treatment. They also constructed ROC curves and used mixed model regression to assess external responsiveness of the utility measures. Thus, it is reasonable to use several analytic approaches to demonstrate the responsiveness of an outcome measure for specific situations. In the present study, four methods are used to evaluate responsiveness in order to allow triangulation of the results: the SRM, Wilcoxon signed rank test, linear mixed model regression, and ROC curve analysis.

1.4.4 Floor and ceiling effects

Floor and ceiling effects refer to clustering of respondents at the lowest or highest possible score, respectively. Terwee et al define the presence of floor or ceiling effects if more than 15% of respondents achieve the lowest or highest score [78]. If these effects are present, patients with the lowest or highest possible scores cannot be distinguished from each other, and reliability is reduced. These effects also limit the ability of the measure to capture changes in scores, which may diminish responsiveness.

1.4.5 Global Rating of Change Scales

As mentioned above, GRCS can be used to qualify a patient's improvement or deterioration over time. GRCS have been used in HRQL research since 1989 [102]. GRCS ask the patient to

assess his/her current health status, recall that status at a previous time-point, and then calculate the difference between the two [103]. The scale has a midpoint of "no change", with increasing degrees of worsening and improvement on either side. In the literature GRCS generally have between 7 and 15 points, and the optimal number of response options is likely between 7 and 11 points [104]. The GRCS can be used as an external criterion for responsiveness studies by classifying patients who have improved or not (e.g. ROC curve analysis) [93, 105], and also to estimate the MCID of HRQL measures by determining the mean change score in patients indicating minimal change on the GRCS [102]. GRCS have many strengths including high face validity and a method to identify a change that is meaningful to the patient [103]. Test-retest reliability and construct validity have been shown to be moderate-high, however responsiveness is variable (SRM=0.2-2.7) [103, 106]. GRCS measures also have several limitations, the most important of which is recall bias, and the difficulty that patients have in judging whether they have changed compared to a past time point [103, 107-109]. GRCS ratings tend to correlate much more with the *current* health state rather than the *change* in change state. In addition, single-item measures are generally less reliable than multi-item measures, calling into question their validity as a 'gold standard' for comparison [109].

1.4.6.1 Use of GRCS in Pain Assessment

The IMMPACT group recommends including a 7-point GRCS as a core outcome measure in clinical trials of adults with chronic pain to enable the participants' assessment of the clinical importance of their improvement or worsening [110, 111]. A pain GRCS is mentioned but not formally recommended by the PedIMMPACT guidelines mainly due to lack of data on this measure in pediatrics [21]. Studies of pain in adults have used 7- or 15- point GRCS to determine the MCID, longitudinal construct validity, and responsiveness by ROC curve analysis of various measures of pain intensity and physical impairment [79, 112, 113].

In the present study, a 5-point GRCS is used as an external criterion for change in pain throughout the course of the study. A 5-point scale was chosen because studies have shown that the limit of human discrimination is approximately 7 response options (range 5-9 options) [114], and the lower end of the range was chosen given the wide spectrum of developmental levels, as has been used in other pediatric studies [115]. The GRCS is used to classify patients as worse, stable or improved with regards to their pain (Methods section 2.4.5.2).

1.5 Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)

This section summarizes the previous studies completed to develop and test the feasibility of the SUPER-KIDZ tool in the JIA population.

1.5.1 SUPER-KIDZ development

Consensus methods involving health professionals and consumers were used to develop the SUPER-KIDZ tool [116]. The measure was developed via a two-phase approach, starting with a Delphi survey to generate domains and items, followed by a consensus conference to determine the final items for inclusion. In the first survey, CARRA members were asked to rate the importance of the 8 PedIMMPACT domains described by McGrath et al [21]. In addition they were asked to provide additional domains they felt were important and to provide specific items within each domain. Based on average importance ratings, all domains were retained except "economic factors". The second Delphi survey asked respondents to rate the importance of several items within each domain, and retention was based on an importance rating of at least 7/10 and level of agreement (free marginal kappa) of at least 0.3 [116].

The second phase of the SUPER-KIDZ development occurred via a 2-day consensus conference based on the results of the Delphi surveys. Interested pediatric rheumatologists, pediatric pain experts, allied health professionals and patient representatives participated in a structured Nominal Group Technique (round-robin voting guided discussion). Consensus conference discussion yielded general agreement on the following domains and items for *inclusion*: 1) pain characteristics (current pain, average pain intensity over past 2 weeks, pain episode duration, pain frequency, pain location), 2) associated symptoms (fatigue frequency), 3) functioning (physical, social, and role) and 4) cognitive and emotional factors (catastrophizing, positive affect, sadness, anger, worry, stressors). Participants agreed to *omit*: pain sensory descriptors, pain aggravating/alleviating factors, pain unpleasantness, comfort goal, global pain treatment satisfaction rating, fatigue, appetite, pain self-efficacy, recent peer group changes or conflicts, and level of independence. Items for the functioning domain were obtained from the PROMIS pediatric pain interference scale [117]. Items for assessing the cognitive dimension of pain (pain catastrophizing) were taken from the three highest loading items on the Pain Catastrophizing Scale for Children [118]. Pain location was captured using a

body outline from von Baeyer and colleagues [119]. The final measures include a self-report version for children aged 4 to 7 years, which includes pain intensity and body location only. The proxy-report version for parents of children aged 4-7 years and the self-report version for youth (8 to 18 years) consist of four domains, three of which also reflect the Melzack multi-dimensional conceptual framework for pain [18, 19]: (1) pain characteristics (sensory-discriminative dimension) (5 items), (2) associated symptoms (1 item), (3) functional interference and catastrophizing (cognitive-evaluative dimension) (7-10 items), and (4) emotional functioning (affective-motivational) (4 items) (Appendix 1).

1.5.2 Feasibility assessment

It is important to determine the acceptability and feasibility of a clinical measure by pre-testing in a small group of users [70, 120]. If an instrument is difficult to use, or puts undue burden on the health care provider or patient, the tool is unlikely to be accepted by medical practitioners or adopted into clinical care.

Several aspects of the feasibility of the SUPER-KIDZ tool were evaluated in the study by Stinson et al [116] in which they piloted the tool in pediatric rheumatology clinics on three mediums: paper, computer and handheld (iPodTouch). A research assistant initially guided the participants through the tool and no difficulty was noted. The feedback from physicians and patients was that the tool was easy to use. The administration time is suitable for use in the clinic setting, taking 3-4 minutes to complete on the computer. There were very few missed responses in their test group (average of 0.1-0.5 questions overall missed on computer medium), indicating that the instructions and questions were easy to understand. From an acceptability point of view, this was evaluated as a comparison of the 3 mediums. Patients least liked the handheld device because of the difficulty with small screen size especially with the body diagram. The computer format was the preferred medium by younger children and physicians; the physicians preferred reviewing the computer printout over the paper version. Overall, it appears that the SUPER-KIDZ tool is developmentally appropriate, acceptable, and easy-to-use by patients and physicians in the clinic setting. The next step in the development of the SUPER-KIDZ measure is to test its measurement properties in the target population.

1.6 Study Rationale and Relevance

As discussed above, pain is the most common and distressing symptom of children and youth with JIA, and is associated with a significant burden of illness and impact on HRQOL. A study by Kimura et al. brought to light the necessity for improvement in pain evaluation and treatment in children with JIA [29]. Thirty-four percent of pediatric rheumatologists in North America indicated that pain was not systematically evaluated or documented in the arthritis patients at their centres. Recently, Lovell et al. recommended assessment of "arthritis-related pain" as a quality of care measure in the treatment of children with JIA [121]. These authors suggest that pain assessment, along with other quality measures, should be routinely implemented in pediatric rheumatology practices as a means of tracking outcomes and generating quality improvement indices.

There is currently no standardized comprehensive approach guiding the clinical assessment of pain in children with JIA presenting to pediatric rheumatologists and other allied health professionals [29, 122]. There is clearly a need for an innovative and developmentally appropriate approach to the assessment of pain in patients with JIA. The SUPER-KIDZ tool represents the first clinically useful pain measure for children and youth with JIA that has been developed according the multi-dimensional pain construct and will fulfill a long-standing gap in clinical care in pediatric rheumatology. It has been found to be feasible for use in the clinic setting, and is well liked by patients and health care providers.

The next step in the development of this tool is to ensure that it is valid for use in patients with JIA. This involves establishing its test-retest reliability, construct validity, responsiveness, and interpretability [71]. Given that the tool is to be used for an evaluative purpose, it is very important to assess test-retest reliability and its responsiveness to change [58], and these objectives are the focus of the current study. This will ensure confident use of the SUPER-KIDZ tool to identify stability and change of the pain experienced by patients with JIA, enabling informed clinical decisions. Construct validity is also critical and is being evaluated concurrently as part of a larger study.

If the measurement properties (reliability, validity, responsiveness, interpretability) are established, there is a plan to implement the SUPER-KIDZ pain tool in all CARRA pediatric rheumatology centres across North America, enabling its use in a large population of patients.

At the patient level, the SUPER-KIDZ pain tool has the potential to improve the assessment and management of pain in JIA patients, thereby reducing unnecessary burden and improving health outcomes. In addition, the computerized measure allows for real-time result read-outs for the clinicians and patients to use immediately during the clinical visit. It is likely that use of the SUPER-KIDZ pain tool will increase patient and physician communication about pain at the clinical encounter, improve patient engagement in decision-making about their pain treatment plan, and ultimately lead to improved child health outcomes (such as HRQL). Another benefit of a standardized clinical pain tool is the potential for integration of pain scores into the patient's electronic medical record.

The SUPER-KIDZ tool will also be incorporated into the CARRAnet database, a longitudinal multicenter multi-disease North American registry for pediatric rheumatology. This will ensure that pain will be routinely assessed in all pediatric rheumatology patients and enable its use as an outcome measure for future CARRA studies.

The purpose of this thesis is to investigate the test-retest reliability and responsiveness of the SUPER-KIDZ tool.

1.7 Objectives and Hypotheses

The objectives of this research study are:

1. To determine the test-retest reliability of the items of the SUPER-KIDZ pain tool when patients' pain is stable.

2. To determine the responsiveness of the items of the SUPER-KIDZ pain tool to changes in patients' pain after intra-articular steroid injection(s).

This study represents an interim analysis of a prospective study. Acceptable performance of the SUPER-KIDZ tool is defined as 75% of the individual SUPER-KIDZ items meeting the a priori criteria specified in the primary and secondary hypotheses (H1-H5) regarding the measurement properties of these items.

The primary hypotheses to be tested are:

- H1: The reliability parameter for each of the SUPER-KIDZ items (ICC for continuous variables and weighted kappa for categorical variables) measured at 2 time points when pain is expected to be stable over the course of one week will be ≥0.80 and the lower bound of the 95% confidence interval will be ≥0.60. This value was chosen based on previous studies [61, 79] showing that pain measures rarely achieve a value of ≥0.90-0.95 typically recommended for individual decision-making [73].

- H2: In patients whose pain is expected to improve after intra-articular steroid injection(s),

    - a) For continuous variables: the point estimate of the SRM for the SUPER-KIDZ items will be low to moderate (SRM ≤ 0.5) at one week post-joint injection, and moderate to high (SRM ≥0.5 – 0.79) at the second week post-joint injection, and the lower bound of the 95% confidence interval will be >0 [65].

    - b) For ordinal variables: the null hypothesis of the Wilcoxon signed rank test (the sum of the ranks of the positive difference scores is equal to the sum of the ranks of the negative difference scores) will be rejected at the second week post-joint injection.

The secondary hypotheses to be tested include:

- H3: At baseline, the lower bound for the ordinal reliability alpha between SUPER-KIDZ items of the same subscale will be 0.70 – 0.95 [78].

- H4: In a linear mixed regression model of SUPER-KIDZ item scores before and after joint injection, there will be a statistically significant improvement in scores 2 weeks after injection compared to baseline (p-value<0.05 for beta coefficient). The magnitude of improvement will be greater at 2 weeks compared with 1 week post-injection [65].

- H5: In the evaluation of responsiveness using an external criterion for improvement, the AUC of the ROC curves will be ≥0.70 [78] and the lower bound of the 95% confidence interval will be >0.50.

2 METHODS

2.1 Study design

This research study is the first phase of a prospective study with repeated measures. The current phase included study set-up (i.e. ethics, funding, training of study personnel, trouble-shooting online measure) and the initial 8 months of subject recruitment. Consecutive JIA patients meeting eligibility criteria were recruited from the Pediatric Rheumatology clinic at the Hospital for Sick Children starting in July 2012. An interim analysis was performed on data collected until the end of February 2013, and is reported herein.

2.2 Research Ethics Board Approval and Funding

Research Ethics Board approval for this study was obtained from the Hospital for Sick Children (REB# 1000031623) and the University of Toronto, Toronto, Ontario. Informed consent, and assent (when appropriate), was obtained from each participant and/or their parent. No significant harms were identified. All patient information was de-identified and linked to a unique study identification number. All study materials were stored in locked cabinets in locked offices of the principal investigator (PI) (J. Stinson). Electronic data was stored in double-password protected documents.

The study was funded through a Discovery Advancement Program Grant from the Canadian Arthritis Network (CAN) (PI J. Stinson, Project code: 11-DAP-14). N. Luca was also partially funded by a Post-Doctoral Fellowship Training Award from CAN.

2.3 Study Population

2.3.1 Inclusion and Exclusion Criteria

There were two main study groups denoted Group A and Group B. Group A participants were scheduled for joint injection(s) and included in both test-retest reliability and responsiveness testing. Group B participants had stable disease management, and were only included in test-retest reliability testing. We included group B, whom we anticipated would have lower levels of pain than group A, in order to avoid potential within-subject variance in scores on repeat

testing due to a "regression to the mean" phenomenon in patients with higher levels of pain before joint injection [123].

2.3.1.1 General inclusion criteria (Groups A and B): Patients a) between the ages of 4-18 years, b) diagnosed with JIA by a rheumatologist (any subtype), c) have self-reported pain in the past week, d) have fluency in English and e) have access to a computer.

2.3.1.2 Additional inclusion criteria: Group A – f) scheduled to undergo a joint injection; Group B – f) no anticipated management changes (test-retest reliability only).

2.3.1.3 Exclusion criteria: Patients were excluded if they had a) a major cognitive or psychiatric disorder (e.g., developmental delay, depression) that may interfere with ability to complete SUPER-KIDZ measure, b) other medical disorders (e.g., Crohn's disease, fibromyalgia) that may contribute to acute or chronic pain, or c) severe vision problems (e.g., cataracts, glaucoma) that may interfere with their ability to see the SUPER-KIDZ measure.

2.3.2 Sample Size

Sample size was calculated for each primary hypothesis and the largest was used.

2.3.2.1 Test-retest reliability

For ICC (2,1) and weighted kappa determination, parameters were set to power 80%, alpha 0.05, one-sided test, and 2 observations per patient. A sample of size of 39 is required to reject the null hypothesis of ICC/kappa of 0.6 if the alternate hypothesis is ICC of 0.8 [124].

2.3.2.2 Responsiveness

For the SRM, two methods were used to calculate the sample size. (1) Using single sample mean: with parameters set to power 80%, alpha 0.05, and one-sided test, the sample size required to reject the null hypothesis of SRM of 0 if alternate hypothesis is SRM=0.5 is 36 [125]. (2) Using confidence interval around SRM: with 95% confidence interval set to ±0.15 and standard deviation set to $1/\sqrt{n}$, a sample size of 44 is required [125].

For the Wilcoxon signed rank test, the possible values for paired differences in the ordinal items (5-point scale) are -4 to +4. Thus the range is 8 and this represents the 95% confidence interval. The SD can be estimated as half the range divided by 1.96 (SD=2.04) [126]. In order

to detect a difference of 1 in the ranks, the effect size is estimated as 1/SD or 1/2.04 = 0.49. For 80% power and alpha 0.05, the required number of pairs is approximately 34 [127].

2.3.2.3 Final sample size

To account for loss-to-follow-up, we aimed to recruit 50 patients to Group A for each of the age groups (young [4-7 years] and older [8-18 years]) (Total Group A=100). To broaden the reliability sample, an additional 30 patients of all ages were added (Group B), for a total of 130 subjects for the complete study.

A planned interim analysis was performed at month 8 of the study when approximately half of the overall sample size had been recruited.

2.4 Study Procedures

2.4.1 SUPER-KIDZ website set-up

The SUPER-KIDZ tool is hosted on a secure website http://superkidzpain.ca. Study personnel create unique study identifications (ID) on the administrative site http://superkidzpain.ca/admin, and this site is also used to verify questionnaire completion and upload subject responses. Prior to study initiation the website was tested by creating dummy patients to ensure that responses uploaded correctly.

2.4.2 Hiring and training of study personnel

A clinical research assistant (CRA) was hired for the study. N. Luca trained the CRA on the theoretical basis for the study and all study procedures including the recruitment process and data management. The administrative staff of the rheumatology clinic also assisted in identifying eligible patients under the direction of N. Luca and the CRA.

2.4.3 Recruitment

The CRA prospectively identified eligible patients on the rheumatology clinic and joint injection lists. The patient's primary health care provider asked the family if they were interested in hearing about the study. If so, the CRA met with the family to explain the study and obtain informed consent and assent (Appendix 2). Recruitment also took place by phone for patients scheduled for joint injection but who were not coming for a clinic visit prior to the

joint injection. In this case, the responsible rheumatologist was asked for permission to send an information letter to the family's home. The CRA then followed up with the family by phone to determine interest, and if so, obtained verbal consent and assent. Participants were compensated with a hospital parking voucher and $5 Tim Horton's gift certificate, as well as community volunteer hours for the older children.

2.4.4 Data collection

2.4.4.1 Initial time point (recruitment)

Demographic and disease-related data were abstracted from the patients' charts at the time of recruitment. These included: the patient's date of birth, gender, diagnosis (JIA subtype), current medications, number of active joints, and physician global assessment (PGA) of disease severity on a 10-cm VAS [128] (Appendix 3).

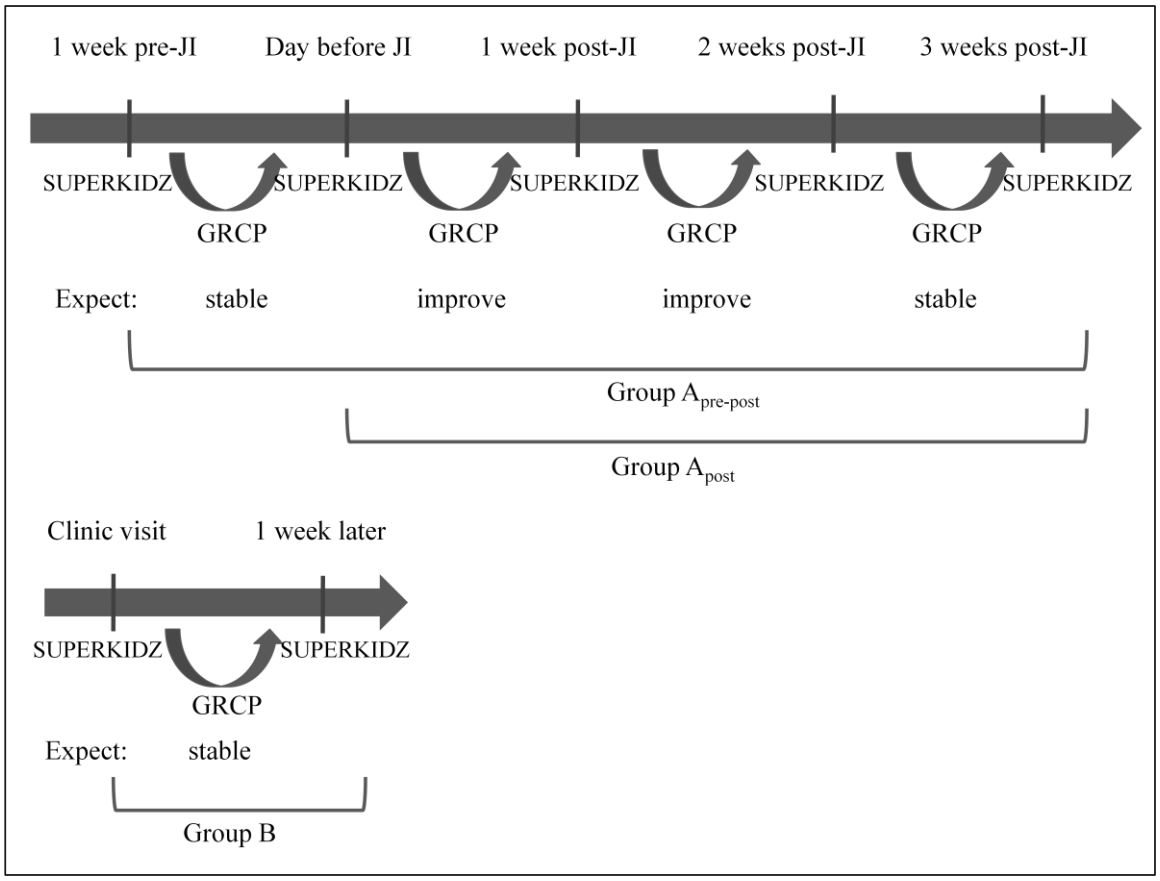
During the recruitment visit, a secure patient account using a unique study ID was created on the SUPER-KIDZ administrative website (http://superkidzpain.ca/admin). The CRA guided the participants in logging into the patient website (http://superkidzpain.ca) and in completing the measure on a laptop computer. For patients recruited by phone, the secure account was set up by the CRA and they completed the questionnaire on their home computer. The CRA verified appropriate completion of the questionnaire on the administrative site.

2.4.4.2 Timeline for repeated measures

A summary of the study procedures for the two groups and the analytic purpose of each time point are summarized in Figure 1 and Appendix 4. Group A patients completed the SUPER-KIDZ tool at five time points: 1 week prior to their joint injection, the day before their joint injection, and then 1, 2, and 3 weeks after their joint injection. The day before the joint injection was used instead of the day of the injection due to the possibility of the patient being anxious on the day of the injection. Group B patients completed the SUPER-KIDZ tool at two time points: in clinic and then one week later (test-retest reliability only). At each time point, participants also completed a global rating of change scale for pain between testing times (i.e. over 1 week) on a five-point ordinal scale, called the Five-Point Global Rating for Change in Pain (GRCP) (see 2.4.5 Measures, below).

For the repeated measures, the CRA contacted patients by either email or telephone to remind them to complete the SUPER-KIDZ questionnaire and the GRCP at the specified time points. A 1-week time period between measurements was chosen as it is long enough to see a difference after joint injection [65] but also short enough that clinical status is unlikely to change in the case of reliability testing [89]. For practical reasons we accepted a minimum of 3 days and maximum of 10 days between the time points.

Figure 1: Schematic timeline of time points for the different study groups



JI=joint injection; GRCP=global rating of change in pain

For Group A, two sets of time points were considered for test-retest reliability: 1) 1-week pre-joint injection and day before joint injection and 2) 2- and 3- weeks post-joint injection. In some cases, Group A patients were missing the 1-week pre-joint injection time point due to recruitment too close to the date of joint injection (less than 3 days). Thus, we defined two groups within Group A according to whether all time points were completed. Participants who completed all time points were labeled Group $A_{pre\text{-}post}$. Those who were missing the 1-week pre-joint injection time point were labeled Group A_{post}. For Group A_{post}, only the data from

weeks 2 and 3 post-joint injection were used for test-retest reliability analysis. For Group B, the initial and 1-week responses were used for test-retest reliability. Responsiveness was assessed in Group A through comparison of both the 1- and 2- week post-joint injection time points to the responses on the day prior to the joint injection (Figure 1).

2.4.5 Measures

2.4.5.1 SUPER-KIDZ Pain Tool

The SUPER-KIDZ pain tool (Appendix 1) is administered online and takes approximately 3-4 minutes to complete [116]. There are three versions: self-report for patients aged 4-7 years, parent-proxy for parents of children aged 4-7 years, and self-report for patients aged 8-18 years. The older child version was used for this study and consists of 20 questions (Table 7).

Table 7: Summary of the 20 SUPER-KIDZ Items

Item	Question	Domain	Melzack Dimension
Q1	How much pain do you have right now?	Pain characteristics	Sensory-discriminative
Q2	If you had pain in the past 7 days, how much did it usually hurt?		
Q3	On how many days did you have pain in the past 7 days?		
Q4	If you had pain in the past 7 days, how long did it usually last?		
Q5	Select all the parts of your body where you have had pain in the past 7 days		
Q6	In the past 7 days, how often have you felt tired?	Assoc'd symptoms	-
Q7	I had trouble sleeping when I had pain	Interference (Physical, social and role functioning)	Evaluative
Q8	It was hard for me to pay attention when I had pain		
Q9	It was hard to stay standing when I had pain		
Q10	It was hard to have fun when I had pain		
Q11	I had trouble doing schoolwork when I had pain		
Q12	It was hard for me to walk 1 block when I had pain		
Q13	It was hard for me to run when I had pain		
Q14	I kept thinking how much I wanted the pain to stop when I had pain		Cognitive
Q15	I was afraid that the pain would get worse when I had pain		
Q16	I felt I couldn't stand it anymore when I had pain		
Q17	How often did you feel sad in the past 7 days?	Emotional functioning	Affective-motivational
Q18	How often did you feel angry in the past 7 days?		
Q19	How often did you feel cheerful in the past 7 days?		
Q20	How often did you feel worried in the past 7 days?		

The SUPER-KIDZ questions reflect the three dimensions described by Melzack [18, 19] and are categorized into four domains: characteristics of pain (sensory-discriminative dimension) (Q1-Q5), associated symptoms (fatigue) (Q6), interference (cognitive-evaluative dimension) (Q7-Q16), and emotional functioning (affective-motivational) (Q17-Q20). Items are answered using 5-, 7- or 11- point scales, except for the body diagram (click on area(s) of pain). Pain intensity is calculated for current pain (Q1) and average pain (Q2) experienced over the previous week on an 11-point numerical rating scale, and given a value from 0 to 10. Pain frequency (Q3) and duration (Q4) are scored on 7- and 6- point ordinal scales respectively, and given numeric values (1-7 and 1-6). The number of body parts (Q5) is counted to indicate total number of painful bodily areas (1-59). The remaining items are scored on a 5-point ordinal scale with values between 0 and 4. A formal scoring algorithm for the SUPER-KIDZ tool has not yet been developed. For the purposes of this study, each item was evaluated separately for test-retest reliability and responsiveness because it is unknown whether the items within the domains are measuring the same aspects of the pain construct.

2.4.5.2 Five-Point GRCP

The five-point GRCP (Appendix 5) is a single question administered either verbally or by email. The question asks about overall change in pain over the previous one week, on a 5-point ordinal scale with the following response options: "much worse", "a little worse", "the same", "a little better" and "much better". A 5-point scale was chosen because studies have shown that the limit of human discrimination is approximately 7 response options (range 5-9) [114], and the lower end of the range was chosen given the paediatric age range. The GRCP was the external criterion for stability and change in this study, as described in previous studies [93, 105, 112]. The definition for stability is a response of "the same". Improvement was assessed at two time points: 1 week and 2 weeks after joint injection. At 1 week after injection, a response of "a little better" or "much better" (i.e. change of 1 or more points on GRCP scale) was considered evidence of improvement. At 2 weeks, since the reference point for the GRCP was the past one-week, specific combinations of response options for the 1- and 2- week time points were defined as evidence of improvement (Table 8).

Table 8: Scenarios for global rating of change in pain (GRCP) responses at 1 week and 2 weeks after joint injection (JI)

GRCP* at 1 week after JI	GRCP* at 2 weeks after JI	Meet criteria for improvement at 2 weeks?
Any worse	Any worse	No
Any worse	Same	No
A little worse	A little better	No
A little worse	Much better	**Yes**
Much worse	A little better	No
Much worse	Same	No
Much worse	Much better	No
Same	Any worse	No
Same	Same	No
Same	Any better	**Yes**
A little better	A little worse	No
A little better	Much worse	No
Much better	A little worse	**Yes**
Much better	Much worse	No
Any better	Same	**Yes**
Any better	Any better	**Yes**

*GRCP is answered relative to the previous 1 week time period
NB. Any better = much or a little better; Any worse = much or a little worse

2.4.6 Data Management

2.4.6.1 Final Dataset

The responses entered on the SUPER-KIDZ website were uploaded into individual excel files containing the repeated measures for each participant. These files were appended and imported into SAS version 9.3 (SAS institute Inc., Carey, NC) to form a complete SAS database containing all repeated observations. Demographic factors, disease-related data, and the GRCP results for each time point were manually input into a separate excel database by one researcher (NL) and the CRA, and this was imported into SAS and linked to the SUPER-KIDZ data using the study ID variable.

The following variables were created from the data (see Appendix 6 Data Dictionary):

1) Age: continuous variable between 4-18 years

2) Gender: male or female

3) JIA subtype: Categorized into 8 categories according to ILAR criteria (Table 1)

4) Number of medications: Interval variable scored between 0 and 8

5) Medications: Categorized into 8 categories: 1) NSAIDs (ibuprofen, naproxen, indomethacin, diclofenac), 2) steroids (prednisone, methylprednisolone), 3) non-biologic DMARD (methotrexate, leflunomide, sulfasalazine), 4) anti-TNFα agents (etanercept, infliximab, adalimumab, golimumab), 5) other biologic agents (abatacept, tocilizumab, anakinra, rituximab), 6) tylenol, 7) narcotic agents (oxycodone, methadone, morphine), 8) calcium and/or vitamin D.

6) Physician Global Assessment (PGA): Continuous variable scored between 0-10.

7) Number of active joints: Interval variable scored between 0-59.

8) SUPER-KIDZ item scores: The items were labeled Q1, Q2, Q3, …,Q20 and scored as follows: Q1 and Q2 from 0-10, Q3 from 1-7, Q4 from 1-6, Q5 from 1-59, and Q6-Q20 from 0-4. For the purposes of analysis Q1, Q2 and Q5 were treated as continuous variables, and the remaining items were treated as ordinal variables. Data manipulation was required for Q4 because the response of '6' was in fact the shortest duration so it became '1' and the other responses moved up by +1. For Q19 the scale was opposite in direction compared to the rest since it was measuring cheerfulness, so the values were reversed in the dataset (i.e. response of '0' became a response of '4').

9) 5-Point GRCP: Coded into categorical variables: 1="much worse", 2= "a little worse", 3="same", 4 = "a little better", 5="much better".

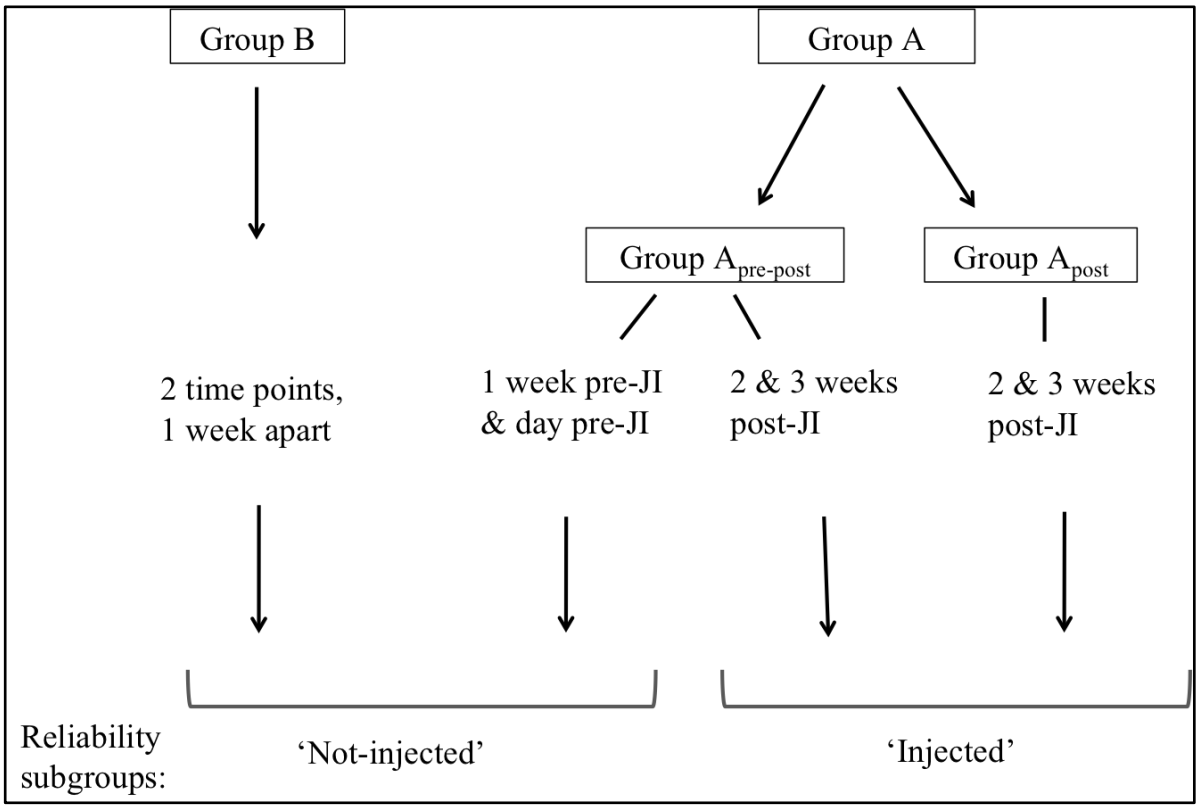
The dataset was examined for extreme observations. In the event of inconsistencies or outliers, the charts and original data were re-reviewed and data corrected as indicated.

2.4.6.2 Defining groups for reliability analysis

For the test-retest reliability analysis, patients who received an injection before the time at which their scores were measured were analyzed separately from patients who had not received an injection. The latter group included patients from both Group $A_{pre\text{-}post}$ (two pre-joint injection time points) and Group B. Thus, the subject categories for reliability testing were: (1)

"Not-injected" group = Group A scores from 1-week pre-joint injection and day before injection combined with Group B scores, and (2) "Injected" group = Group A scores from 2- and 3- weeks post-joint injection (Figure 2). A subject was considered "stable" if he/she responded that his/her pain was the "same" on the 5-point GRCP.

Figure 2: Combining data from study groups for reliability analysis



JI=joint injection

2.5 Data Analysis

Statistical analyses were performed using SAS version 9.3 (SAS Institute Inc., Carey, NC) for the majority of analyses, and Stata version 11.2 (College Station, Texas) for the ROC curve analyses. Statistical significance was defined as a p-value of <0.05. An adjustment for multiple comparisons was not performed based on the fact that we are not testing the hypothesis that all null hypotheses are true simultaneously [129].

2.5.1 Description of Study Cohort

The characteristics of the study cohort were described with regards to age, gender, JIA subtype, medication use, number of active joints and PGA. Descriptive statistics are reported as mean

(±SD) for continuous variables and proportions for categorical variables. If the variable was not normally distributed, the median (inter-quartile range [IQR]) was reported. The Shapiro-Wilk test was used to assess normality.

Groups $A_{pre\text{-}post}$, A_{post} and B were compared according to demographic and disease-related features. For continuous variables, the student's T test (if normal) or Wilcoxon Rank test (if non-normal) was used. For categorical variables, the Chi-square test or Fisher's exact test (if <5 observations per category) was used.

2.5.2 Distribution of baseline scores

The frequency of baseline scores for each item of the SUPER-KIDZ tool was calculated for the whole sample. In addition, the mean score, median score, SD, and IQR were calculated for each item. The distribution of the baseline scores for each item was plotted graphically and examined for floor or ceiling effects resulting from clustering of the data at the low or high ends of the scales. The item scores were compared amongst the 3 study groups (Groups $A_{pre\text{-}post}$, A_{post} and B) using the Wilcoxon Rank Sum test to see if there were any differences.

2.5.3 Internal consistency of SUPER-KIDZ subscales

Measures of internal consistency were determined for the following SUPER-KIDZ subscales: pain characteristics (5 items, Q1-Q5), interference (10 items, Q7-Q16), and emotional functioning (4 items, Q17-Q20).

The ordinal reliability coefficient α was calculated for each SUPERK-KIDZ domain using polychoric inter-item correlations according to the formula [75]:

$$\text{Ordinal reliability } \alpha = k*r_{avg}\,[1+(k-1)*r_{avg}] \qquad \text{(Equation 9)}$$

Where k=number of items and r_{avg} is the average polychoric inter-item correlation. Ordinal reliability α was also calculated after consecutively removing each of the items in the subscale to see whether there was increase in the alpha parameter.

For each individual item, the ITC was calculated as the correlation between the score of the item under study and the sum of the remaining items in the corresponding SUPER-KIDZ subscale [72].

2.5.4 Test-Retest Reliability

The primary test-retest reliability analysis was performed for each item of the SUPER-KIDZ measure using data from all participants. A secondary analysis was limited to the subgroup of subjects who indicated no change in pain after 1 week on their GRCP.

An ICC (2,1) or weighted kappa of at least 0.80 was considered to be a minimum standard for reliability and the ability to interpret questionnaire scores in individual patients. This is less than the 0.90-0.95 range typically cited because pain is known to fluctuate considerably, and the reliability coefficients of other generic pain intensity measures has ranged between 0.63 and 0.90 [61, 79].

2.5.4.1 Distribution of GRCP responses

The distribution of the GRCP responses was determined for both the "not-injected" and "injected" reliability subgroups. The number and proportion of study subjects whose pain was stable over the one-week period (answer = "the same") was determined for each subgroup.

2.5.4.2 Continuous variables: Intra-class correlation coefficient

The Shrout and Fleiss ICC (2,1) model was chosen for the analysis [77] because the testing framework of this study is only one of many possible ways to assess test-retest reliability. A two-way random-effects repeated-measures ANOVA was performed for each SUPER-KIDZ item. The sums of squares of the variances were determined, including: total sums of squares (SS_{TOTAL}), between-subjects mean square (BMS), between-trials mean square (RMS) and error mean square (EMS). The ICC (2,1) for each item was calculated according to Equation 3 (page 15) [77]. A 95% confidence interval around the ICC was calculated by the method of McGraw and Wong[130] according to the formulas for ICC (A,1) using the mean square results from the corresponding ANOVA:

$$\text{Lower limit} = (n(BMS - F^{*}EMS)) / (F^{*}[k*RMS + (kn-k-n)EMS] + nBMS) \quad \text{(Equation 10a)}$$

$$\text{Upper limit} = (n(F_{*}BMS - EMS)) / (k*RMS + (kn-k-n)EMS + nF_{*}BMS) \quad \text{(Equation 10b)}$$

Where F^{*} (F_{*}) denotes the (1- ½ α) x 100^{th} percentile of the F distribution with n-1 (v) numerator degrees of freedom and v (n-1) denominator degrees of freedom.

2.5.4.3 Continuous variables: Error parameters and minimal detectable change

For each continuous item, the SEM for an individual measurement was calculated as follows [76]:

$$SEM = SD \sqrt{(1-r_{xx})} \quad \text{(Equation 6)}$$

where SD is the standard deviation of the observed test scores at baseline (estimated from the corresponding ANOVA as $SD=\sqrt{SS_{TOTAL}/[n-1]}$), and r_{xx} is the reliability coefficient (ICC [2,1]) for that measurement. The standard error of the difference scores (SE_{diff}) was determined by multiplying by the SEM by $\sqrt{2}$ [81],

$$SE_{diff} = SD \sqrt{2}(1-r_{xx}) \quad \text{(Equation 7)}$$

The SE_{diff} was used to calculate the MDC at the 95% confidence level for each continuous item of the SUPER-KIDZ tool using the formula [76]:

$$MDC_{95} = SE_{diff} \times 1.96 \quad \text{(Equation 8b)}$$

2.5.4.4 Ordinal variables: Weighted Cohen's Kappa

Percent agreement was calculated as the number of exact agreements divided by the number of possible agreements. Cohen's weighted kappa was calculated as in Equation 5 [74], and reported with 95% confidence intervals.

$$\kappa = 1 - ([\Sigma w_{ij} \times P_{oij}] / [\Sigma w_{ij} \times P_{eij}]) \quad \text{(Equation 5)}$$

Incremental quadratic Fleiss-Cohen weights were assigned by assuming that the ordinal scale is a continuum with equal intervals. The quadratic weights were calculated as follows [131]:

$$w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_C - C_1)^2} \quad \text{(Equation 11)}$$

where C_i is the score for column i, C_j is the score for column j, and c is the number of categories or columns. For 5 response options (0, 1, 2, 3, 4), the weights are as follows: w_{01}=0.94, w_{02}=0.75, w_{03}=0.44, w_{04}=0, w_{12}=0.94, w_{13}=0.75, w_{14}=0.44, w_{23}=0.96, w_{24}=0.75, w_{34}=0.94.

2.5.5 Responsiveness

2.5.5.1 Types of change

According to the responsiveness taxonomy described above (section 1.4.2) [84], we looked at two types of change over the study period. For the primary analysis, we hypothesized that patients would likely, on average, have improvement in pain after their joint injection based on the known efficacy of joint injections [8, 132]. Therefore we compared pre-injection scores with 1- and 2- week post-injection scores. In this case, the change construct being measured was overall change in pain at the group level ('who'), within-person change over time ('which'), and observed change in the population undergoing joint injection ('what'), or internal responsiveness. In a secondary analysis, we evaluated external responsiveness using the external criterion of improvement based on the GRCP response, and the change construct ('what') being measured was the observed change in population deemed to have improved.

2.5.5.2 Distribution of GRCP responses

The distribution of GRCP responses was plotted for the 1- and 2- week post-joint injection responsiveness time points. The patients who improved at 1-week post-injection were identified as those responding "a little better" or "much better". The patients meeting the criteria for improvement at 2 weeks were identified as described in Table 8.

2.5.5.3 Distribution of SUPER-KIDZ item scores over time

The initial step in the responsiveness analysis was to examine the scores from the three relevant time points (one day pre-injection, 1-week post-injection and 2-week post-injection) for each SUPER-KIDZ item. The data was plotted using box plots to show the mean, median, IQR, and total range at each time point. In addition, the distribution of the change scores at 2-weeks post-injection was plotted for subjects reporting improvement and those not improved.

2.5.5.4 Continuous variables: Standardized response mean

The SRM was calculated for the continuous SUPER-KIDZ items at both the 1- and 2- week post-injection time points using the mean observed change scores and SD of the difference scores according to the equation [92]:

$$SRM = \text{mean}\,(test_2\text{-}test_1) / SD_{diff} \qquad \text{(Equation 12)}$$

Confidence intervals were constructed for the SRM under the assumption that the difference scores, similar in structure to the standardized z statistic, were normally distributed. Therefore the SRM distribution could be approximated by a standard normal distribution with a mean of zero and SD of $1/\sqrt{n}$ where n = sample size [133, 134]. Confidence intervals (95%) were calculated using the z score for 95% confidence:

$$95\%\ CI = \pm 1.96*(1/\sqrt{n}) \qquad \text{(Equation 13)}$$

2.5.5.5 All variables: Wilcoxon signed rank test

For all SUPER-KIDZ items, the Wilcoxon signed rank was performed to determine whether there was a significant improvement in scores compared with baseline at both 1- and 2- weeks post-injection. Subtracting the sets of scores created difference variables for each analysis. The null hypothesis is that the sum of the ranks of the positive difference scores is equal to the sum of the ranks of the negative difference scores. The alternative hypothesis is that the sum of the ranks of the positive difference scores is not equal to the sum of the ranks of the negative difference scores. The S statistic and p-value are reported for each item.

2.5.5.6 Regression analysis

In the regression-based approach, a linear mixed model was fit for each SUPER-KIDZ item with the item scores as the dependent variable and time as the independent variable. Data from the day pre-injection, 1-week post-injection, and 2-week post-injection time points were included in the analysis. The model was run using several different possible variance correlation structures (i.e. compound symmetry, autoregressive etc), and the model of best fit was determined based on the lowest Akaike's Information Criterion (AIC). The Omnibus F test and p-value are reported for the overall model. For each combination of time points, the beta coefficient, standard error, 95% confidence interval, and p-value are reported.

The key assumptions for a linear mixed model were verified. Normality of the residuals was assessed graphically via a quantile-quantile plot. If normality was violated, a transformation of the data was attempted to see if this improved the residual distribution. Potential influential outliers were identified by examining the Cook's D parameter, which combines information on the leverage and residual of an observation. Cook's D measures the change in the model parameter estimates when the observation in question is deleted, and the general cut-off for a

potential influential observation is >4/n [135]. If identified, the model was re-run without the influential outlier(s) to see if the regression coefficients and/or p-values changed substantially. Homoscedasticity of the variances, and therefore appropriateness of the chosen correlation structure, was assessed by plotting the predicted values versus the studentized residuals and looking for a consistent pattern of data points around the regression line.

2.5.5.7 Receiver operating characteristic curve analysis

A ROC curve was constructed as a method to evaluate external responsiveness of the SUPER-KIDZ items at 2 weeks post-injection. The external criterion was improvement at 2 weeks according to GRCP response as defined in Table 8. We considered change scores of -1 to +3 for questions 1 and 2, -1 to +4 for question 5, and -1 to +2 for the other SUPER-KIDZ items. Each change score was compared with the external criterion (GRCP response) to see whether it corresponded to an improvement in the patient.

For each cut-off point, the sensitivity was calculated as:

$$\text{Sensitivity} = \text{True positive} / (\text{True positive} + \text{False negative}) \quad \text{(Equation 14a)}$$

$$\text{Sensitivity} = \frac{\#\text{ patients with change} \geq \text{cut-off who improved}}{\text{total \# patients who improved}} \quad \text{(Equation 14b)}$$

and the specificity was calculated as:

$$\text{Specificity} = \text{True negative} / (\text{True negative} + \text{False positive}) \quad \text{(Equation 15a)}$$

$$\text{Specificity} = \frac{\#\text{ patients with change} < \text{cut-off who did not improve}}{\text{total \# patients who did not improve}} \quad \text{(Equation 15b)}$$

Sensitivity (y-axis) is plotted against 1-specificity (x-axis) to create the ROC curves.

The accuracy of the cut-off point was calculated as:

$$\text{Accuracy} = \frac{\#\text{ patients with change} \geq \text{cut-off correctly classified}}{\text{total \# patients with change} \geq \text{cut-off}} \quad \text{(Equation 16)}$$

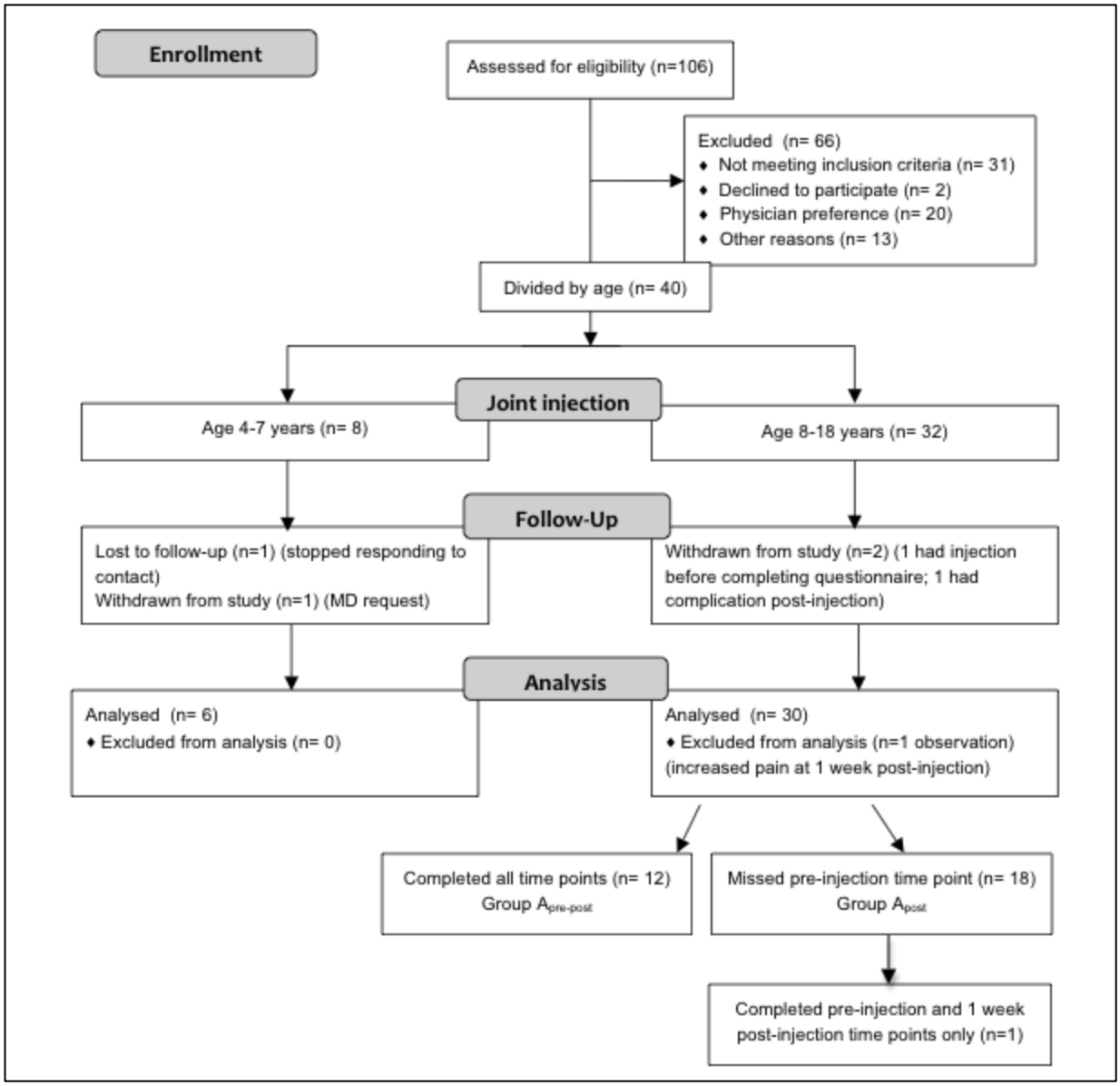
The AUC and its standard error were calculated for the ROC curves to represent the responsiveness of the SUPER-KIDZ item being evaluated.

3 RESULTS

3.1 Description of study cohort

At the time of interim analysis (Feb 23, 2013), 106 subjects had been scheduled for joint injection at the hospital and were screened for eligibility into Group A (Figure 3).

Figure 3: Flow diagram for patients undergoing joint injection considered for inclusion into Group A



Of these, 66 patients were excluded because they were less than 4 years old (n=14), not reporting pain (n=8), met one of the exclusion criteria (n=9), declined participation (n=2), felt not to be appropriate according to responsible physician (n=20), or other reason (n=13). Of the remaining 40 patients, 32 were aged 8-18 years and 8 were 4-7 years of age. In the older age group, 2 patients were enrolled but then were withdrawn from the study: one underwent joint injection before completing the questionnaire, and the other had a severe complication after the joint injection. Thus, 30 subjects completed the study time points in Group A. Within Group A, 12 (40%) completed the pre- and post- joint injection time points (Group $A_{pre\text{-}post}$) and 18 (60%) completed the post-injection time points only (Group A_{post}). One patient in Group A_{post} only completed 2 time points - the day before and 1 week after injection - and was included in the 1-week responsiveness analysis only. Consecutive patients with self-reported pain and stable management were enrolled into Group B (n=21), for a total of 51 patients. Enrollment in the younger age group (aged 4-7 years) is ongoing and currently there are 12 patients enrolled (n=6 in joint injection group, n=6 in stable group).

In the overall sample, the median age was 13.9 years (IQR=11.5-16.2), and 40 (78%) were female. The most common JIA subtypes were: RF negative polyarticular (29%) and persistent and extended oligoarticular (18% each). At baseline, the median joint count was 3 (IQR=1-5) and median PGA score was 2.5 cm (IQR=2.5-4.0) (Table 9).

Median age, number of active joints, PGA and number of medications were similar across the study groups based on Wilcoxon sum rank tests. Group B had no patients with the persistent oligoarticular subtype (Fisher's $P=7.97x10^{-4}$, p=0.003). Group $A_{pre\text{-}post}$ had no extended oligoarticular patients, however the difference in proportions was not statistically significant (Fisher's P=0.02, p=0.18). The proportion of other JIA subtypes, the gender distribution, and medication types were similar between the three groups based on Fisher's exact test.

Table 9: Demographic and clinical characteristics of the whole study sample and subgroups

Characteristic	Whole Sample (n=51)	Group $A_{pre\text{-}post}$ (n=12)	Group A_{post} (n=18*)	Group B (n=21)
Age (years) (median, IQR)	13.9 (11.5-16.2)	13.1 (11.3-15.6)	14.9 (10.2-16.1)	14.4 (11.6-17.1)
Gender (female) (n, %)	40 (78%)	10 (83%)	14 (78%)	16 (76%)
JIA subtype (n, %)				
Oligo persistent	9 (18%)	5 (42%)	4 (22%)	0 (0%)
Oligo extended	9 (18%)	0 (0%)	4 (22%)	5 (24%)
Poly RF negative	15 (29%)	4 (33%)	4 (22%)	7 (33%)
Poly RF positive	6 (12%)	1 (8%)	2 (11%)	3 (14%)
Systemic	3 (6%)	0 (0%)	2 (11%)	1 (5%)
Enthesitis-related	3 (6%)	1 (8%)	0 (0%)	2 (10%)
Psoriatic	4 (8%)	1 (8%)	2 (11%)	1 (5%)
Undifferentiated	2 (4%)	0 (0%)	0 (0%)	2 (10%)
Number of meds (median, IQR)	2 (1-2)	2 (1-3)	2 (1-2)	2 (1-2)
Medications (n, %)				
NSAIDs	28 (55%)	7 (58%)	11 (61%)	10 (48%)
Steroid	7 (14%)	2 (17%)	2 (11%)	3 (14%)
Non-biologic DMARDs	30 (59%)	8 (67%)	8 (44%)	14 (67%)
Anti-TNFα agents	13 (25%)	2 (17%)	7 (39%)	4 (19%)
Other biologic agents	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Tylenol	10 (20%)	2 (17%)	3 (17%)	5 (24%)
Narcotic agents	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Calcium/vitamin D	9 (18%)	3 (25%)	3 (17%)	3 (14%)
Number of active joints (median, IQR)	3 (1-5)	3 (2-4)	3 (1-4)	3 (1-5)
PGA (cm) (median, IQR)	2.5 (1.5-4.0)	2.2 (1.2-3.3)	3.0 (2-3)	2.0 (1-5)

Group $A_{pre\text{-}post}$: Patients undergoing joint injection (JI) completing all time points; Group A_{post}: Patients undergoing JI, missing week pre-JI time point (*includes 1 patient with only 2 time points); Group B: Stable patients used for reliability only.

3.2 Distribution of baseline scores

Recalling the 20 SUPER-KIDZ questions (Table 7, re-presented below), 19 of the items were not normally distributed, and Q2 was just marginally normal (Shapiro Wilk W=0.96, p=0.08). Thus, non-parametric methods were used for all subsequent comparison analyses.

Table 7**:** Description of the 20 SUPER-KIDZ Items

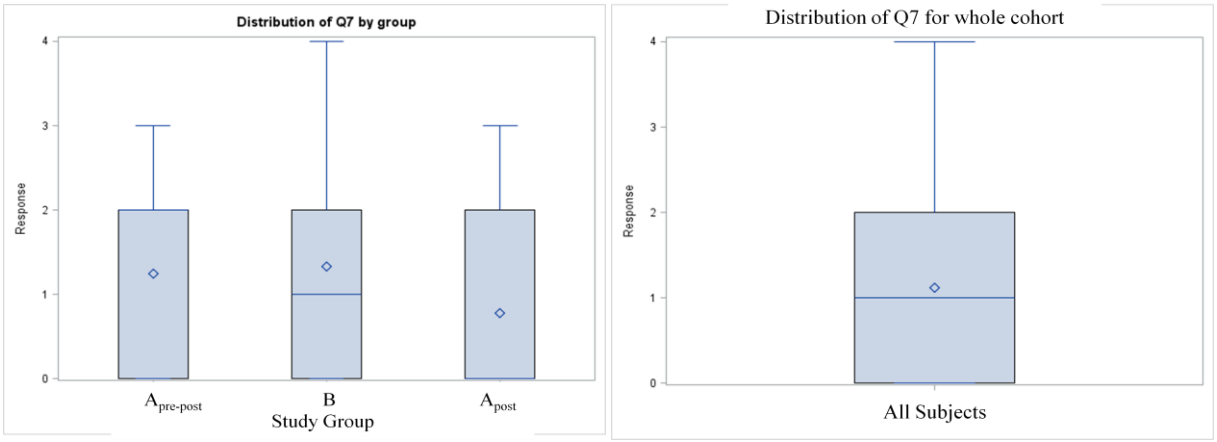
Item	**Question**	**Score range**	**Domain**
Q1	How much pain do you have right now?	0-10	Pain characteristics
Q2	If you had pain in the past 7 days, how much did it usually hurt?	0-10	
Q3	On how many days did you have pain in the past 7 days?	1-7	
Q4	If you had pain in the past 7 days, how long did it usually last?	1-6	
Q5	Select all the parts of your body where you have had pain in the past 7 days	1-59	
Q6	In the past 7 days, how often have you felt tired?	0-4	Associated symptoms
Q7	I had trouble sleeping when I had pain	0-4	Interference (Physical, social and role functioning, cognitive behaviours)
Q8	It was hard for me to pay attention when I had pain	0-4	
Q9	It was hard to stay standing when I had pain	0-4	
Q10	It was hard to have fun when I had pain	0-4	
Q11	I had trouble doing schoolwork when I had pain	0-4	
Q12	It was hard for me to walk one block when I had pain	0-4	
Q13	It was hard for me to run when I had pain	0-4	
Q14	I kept thinking how much I wanted the pain to stop when I had pain	0-4	
Q15	I was afraid that the pain would get worse when I had pain	0-4	
Q16	I felt I couldn't stand it anymore when I had pain	0-4	
Q17	How often did you feel sad in the past 7 days?	0-4	Emotional functioning
Q18	How often did you feel angry in the past 7 days?	0-4	
Q19	How often did you feel cheerful in the past 7 days?	0-4	
Q20	How often did you feel worried in the past 7 days?	0-4	

The median scores for current pain intensity (Q1) and average pain intensity over past week (Q2) were 4.0 (IQR=2.0-6.0) and 5.0 (IQR=3.0=7.0), respectively. Of the items with 5 response options, the median score for Q14 was 3.0 (IQR=2.0-4.0), while the medians of the other items were either 1.0 or 2.0. Of note, no subjects chose the highest response option for Q2, Q17 and Q18 (see Appendix 7, table summary of baseline scores).

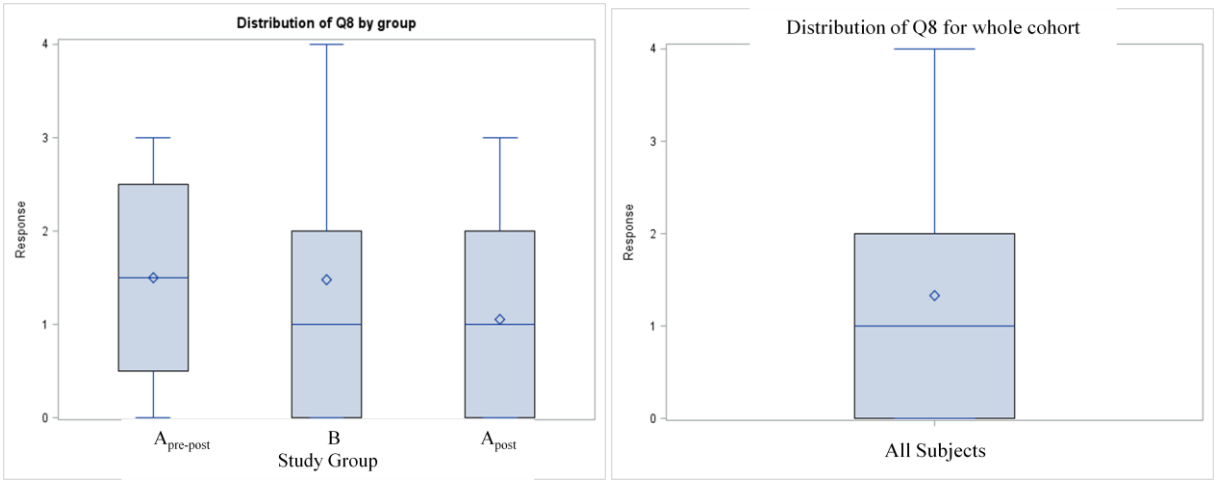
The most striking floor effects (Figure 4) were identified for Q7 (21/51 [41%] gave lowest response), Q8 (16/52 [31%] gave lowest response), Q11 (20/51 [39%] patients gave lowest response), Q17 (15/51 [29%] gave lowest response), and Q18 (19/51 [37%] patients gave lowest response). In contrast, ceiling effects (Figure 5) were seen in Q3 and Q4, where 17 (33%) subjects gave the highest response option for each, and also in Q13 and Q14, where 14 (27%) and 13 (25%) subjects clustered in the highest response categories respectively (see Appendix 7 for figures for other items).

Figure 4: SUPER-KIDZ items with floor effects (Q7, Q8, Q11, Q17, Q18)

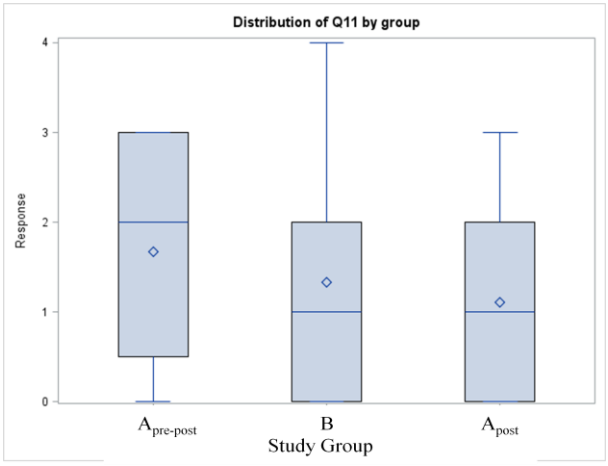
Question 7: Difficulty sleeping

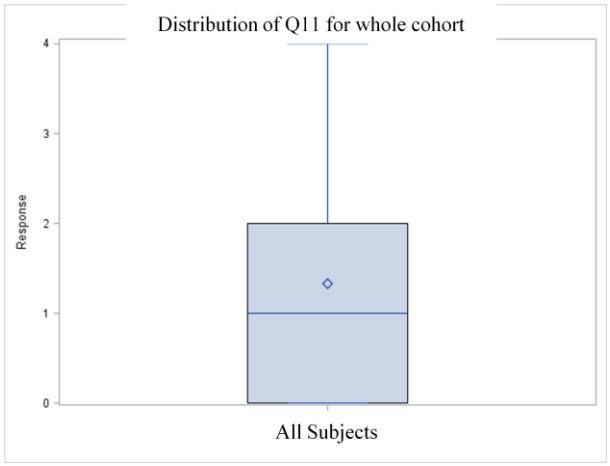


Question 8: Difficulty paying attention

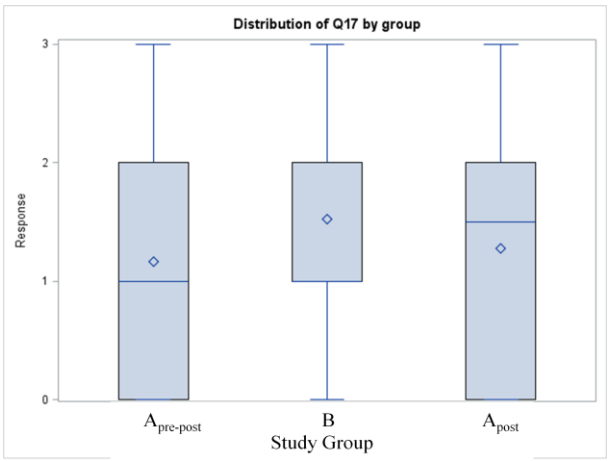


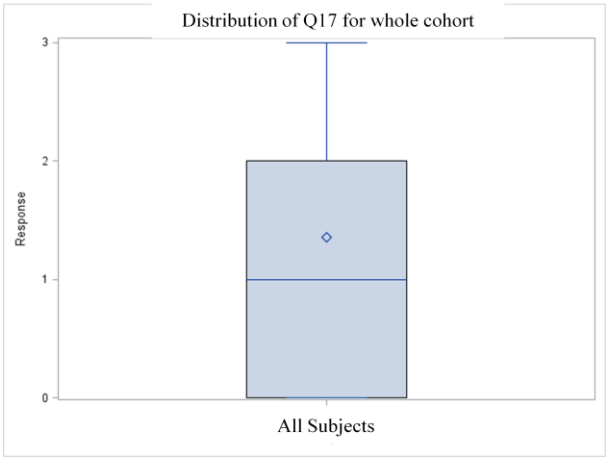
Question 11: Difficulty doing schoolwork



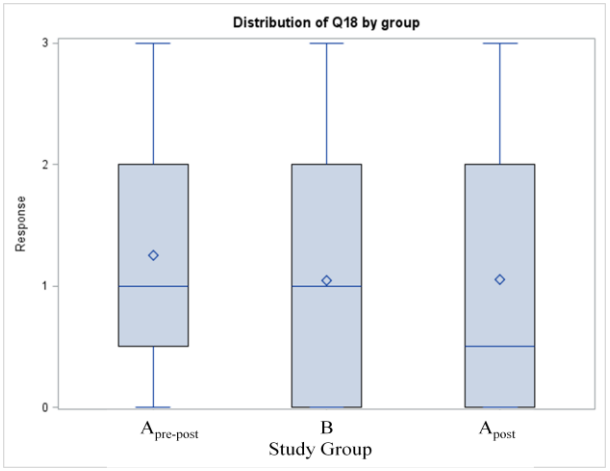


Question 17: Felt sad





Question 18: Felt angry



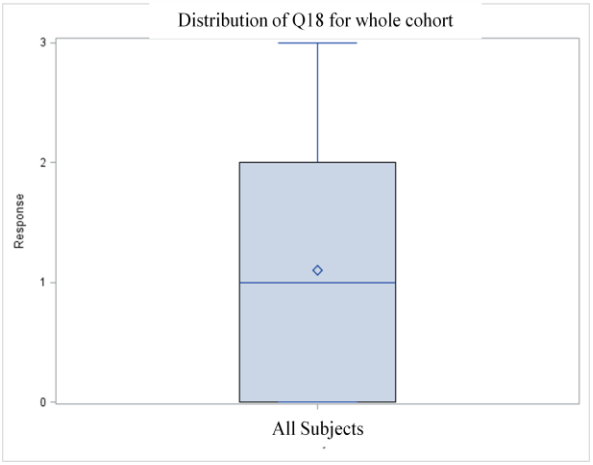
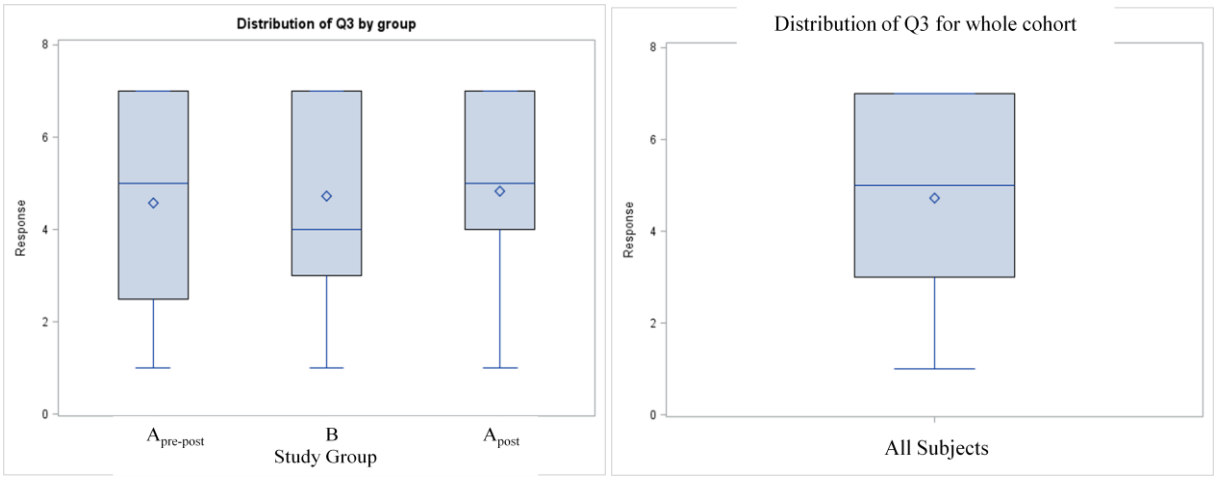
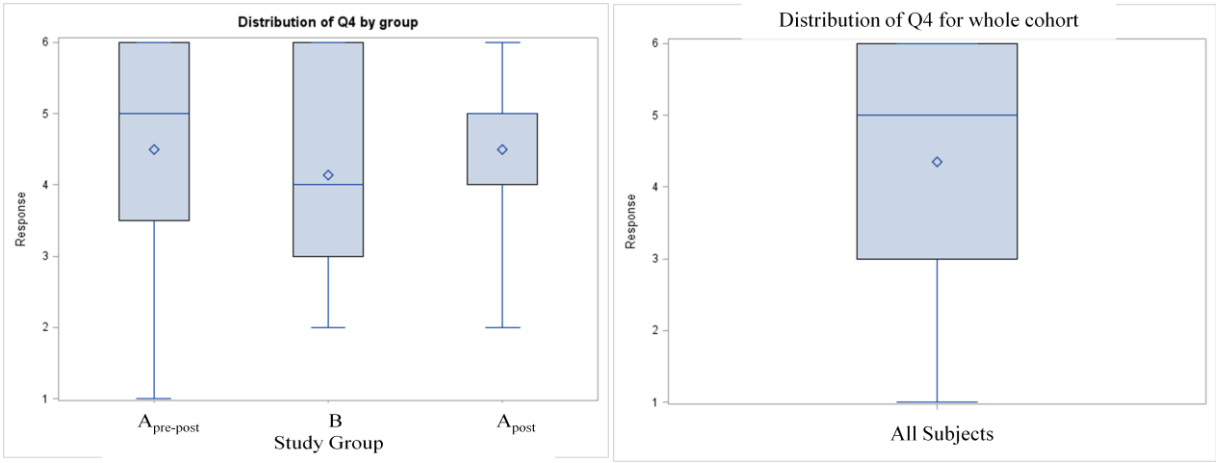


Figure 5: SUPER-KIDZ items with ceiling effects (Q3, Q4, Q13, Q14)

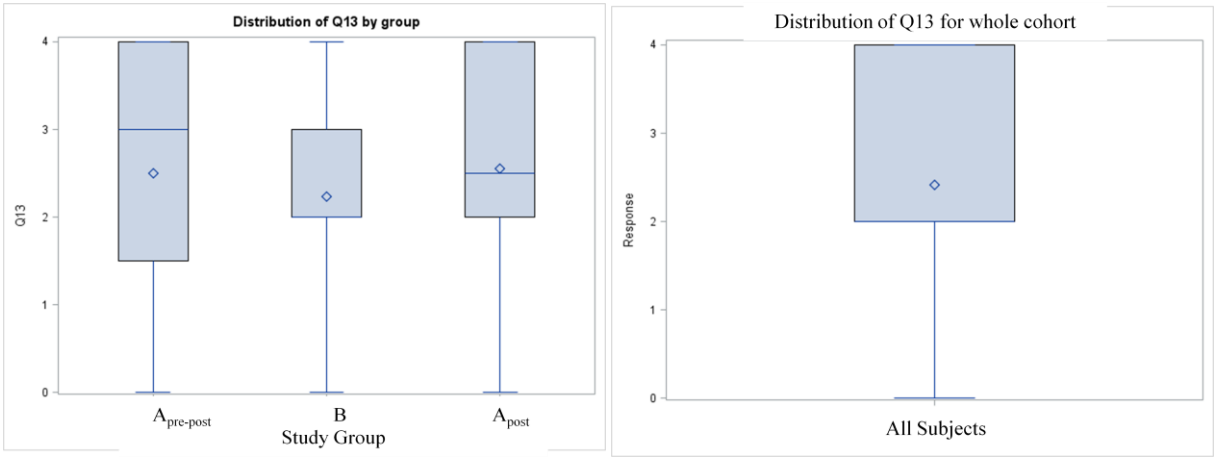
Question 3: Frequency of pain



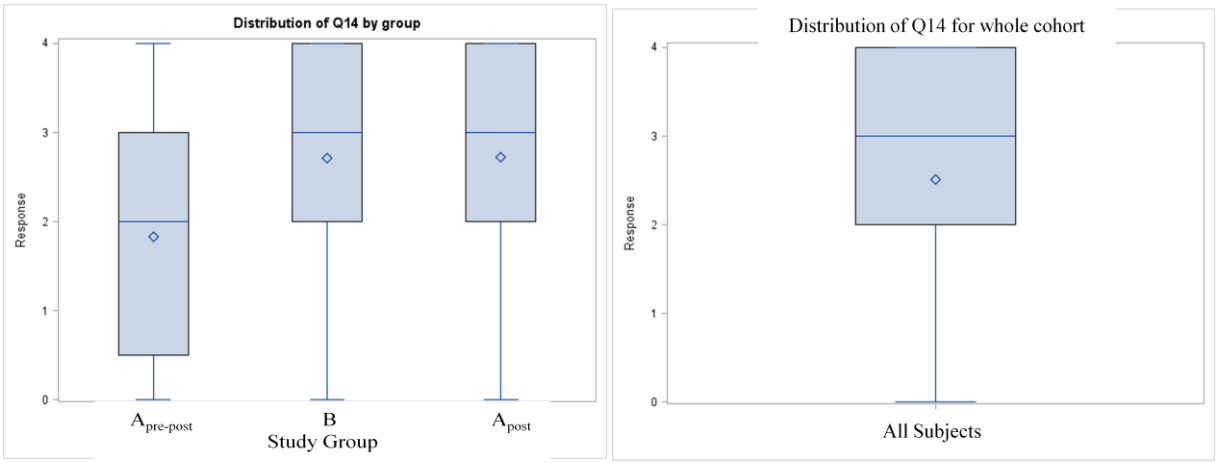
Question 4: Pain duration



Question 13: Difficulty running



Question 14: Kept thinking how much wanted pain to stop



As seen in Table 10, the item scores were similar amongst the study groups at baseline based on the Wilcoxon Rank Sum test (all p values >0.17).

Table 10: Comparison of medians and interquartile ranges (IQR) of baseline item scores amongst the study groups

Item (response range)	**All (n=51)** Median (IQR)	**Group $A_{pre-post}$ (n=12)** Median (IQR)	**Group A_{post} (n=18)** Median (IQR)	**Group B (n=21)** Median (IQR)	**p value***
Q1 (0-10)	4.0 (2.0-6.0)	5.5 (2.0-7.0)	4.0 (3.0-5.0)	4.0 (1.0-6.0)	0.58
Q2 (0-10)	5.0 (3.0-7.0)	5.0 (3.0-7.0)	4.0 (3.0-6.0)	5.0 (3.0-7.0)	0.83
Q3 (1-7)	5.0 (3.0-7.0)	5.0 (2.5-7.0)	5.0 (4.0-6.0)	4.0 (3.0-7.0)	0.99
Q4 (1-6)	5.0 (3.0-6.0)	5.0 (3.5-6.0)	5.0 (4.0-5.0)	4.0 (3.0-6.0)	0.77
Q5 (1-59)	5.0 (2.0-9.0)	4.5 (2.0-7.0)	4.0 (2.0-8.0)	7.0 (2.0-12)	0.47
Q6 (0-4)	2.0 (2.0-3.0)	2.0 (2.0-2.5)	2.0 (2.0-3.0)	3.0 (2.0-3.0)	0.24
Q7 (0-4)	1.0 (0-2.0)	2.0 (0-2.0)	0 (0-2.0)	1.0 (0-2.0)	0.28
Q8 (0-4)	1.0 (0-2.0)	1.5 (0.5-2.5)	1.0 (0-2.0)	1.0 (0-2.0)	0.39
Q9 (0-4)	2.0 (1.0-3.0)	2.0 (1.0-2.5)	2.0 (1.0-3.0)	2.0 (1.0-3.0)	0.85
Q10 (0-4)	2.0 (1.0-3.0)	1.5 (0.5-3.0)	2.0 (1.0-3.0)	2.0 (1.0-2.0)	0.95
Q11 (0-4)	1.0 (0-2.0)	2.0 (0.5-3.0)	0 (0-2.0)	1.0 (0-2.0)	0.40
Q12 (0-4)	2.0 (1.0-2.0)	2.0 (1.0-3.0)	2.0 (2.0-2.0)	2.0 (1.0-2.0)	0.49
Q13 (0-4)	2.0 (2.0-4.0)	3.0 (1.5-4.0)	3.0 (2.0-4.0)	2.0 (2.0-3.0)	0.47
Q14 (0-4)	3.0 (2.0-4.0)	2.0 (0.5-3.0)	3.0 (2.0-3.0)	3.0 (2.0-4.0)	0.17
Q15 (0-4)	2.0 (1.0-3.0)	2.0 (0.5-2.5)	2.0 (1.0-2.0)	2.0 (2.0-3.0)	0.54
Q16 (0-4)	1.0 (0-2.0)	1.0 (0-1.5)	1.0 (1.0-2.0)	2.0 (1.0-3.0)	0.45
Q17 (0-4)	1.0 (0-2.0)	1.0 (0-2.0)	2.0 (0-2.0)	1.0 (1.0-2.0)	0.67
Q18 (0-4)	1.0 (0-2.0)	1.0 (0.5-2.0)	1.0 (0-2.0)	1.0 (0-2.0)	0.84
Q19 (0-4)	2.0 (1.0-2.0)	2.0 (1.0-2.0)	2.0 (1.0-2.0)	2.0 (1.0-2.0)	0.64
Q20 (0-4)	2.0 (0-2.0)	2.0 (0.5-2.5)	2.0 (0-2.0)	2.0 (1.0-3.0)	0.48

* Wilcoxon Rank Sum test

3.3 Internal consistency of SUPER-KIDZ subscales

Internal consistency for each item measured by the ITC was generally greater than 0.50, except for Q5 (number of painful body parts) and Q19 (feeling cheerful), which had more borderline ITCs of 0.40 and 0.25 respectively. Their removal resulted in an increased ordinal reliability α for the respective domain (Table 11). The ordinal reliability α parameters for the SUPER-KIDZ subscales are moderate-high in magnitude (α=0.73-0.92) and all met the hypothesized criteria in H3. The largest value (α=0.92) corresponded to the domain with the most items (interference) while the 4-item emotional domain had the smallest value (α=0.73).

Table 11: Internal consistency measures of items and domains of SUPER-KIDZ questionnaire measured at baseline (raw variables) (n=51)

Domain	**Item**	**Question**	**Item-total correlation (ITC)**	**Ordinal reliability α for subscale**	**Ordinal reliability α if item removed**
Pain characteristics	Q1	Current pain	0.61	0.87	0.83
	Q2	Avg pain past wk	0.71		0.82
	Q3	Pain frequency	0.55		0.83
	Q4	Pain duration	0.51		0.86
	Q5	# Painful locations	0.40		0.89
Interference	Q7	Sleeping	0.67	0.92	0.91
	Q8	Paying attention	0.67		0.91
	Q9	Standing	0.72		0.90
	Q10	Having fun	0.83		0.89
	Q11	Schoolwork	0.54		0.92
	Q12	Walk 1 block	0.64		0.91
	Q13	Running	0.58		0.91
	Q14	Want pain to stop	0.54		0.92
	Q15	Afraid get worse	0.54		0.92
	Q16	Can't stand it	0.65		0.91
Emotional functioning	Q17	Feel sad	0.61	0.73	0.59
	Q18	Feel angry	0.58		0.59
	Q19	Feel cheerful	0.25		0.80
	Q20	Feel worried	0.49		0.68

3.4 Test-retest reliability

3.4.1 Missing data and final sample size for test-retest reliability analyses

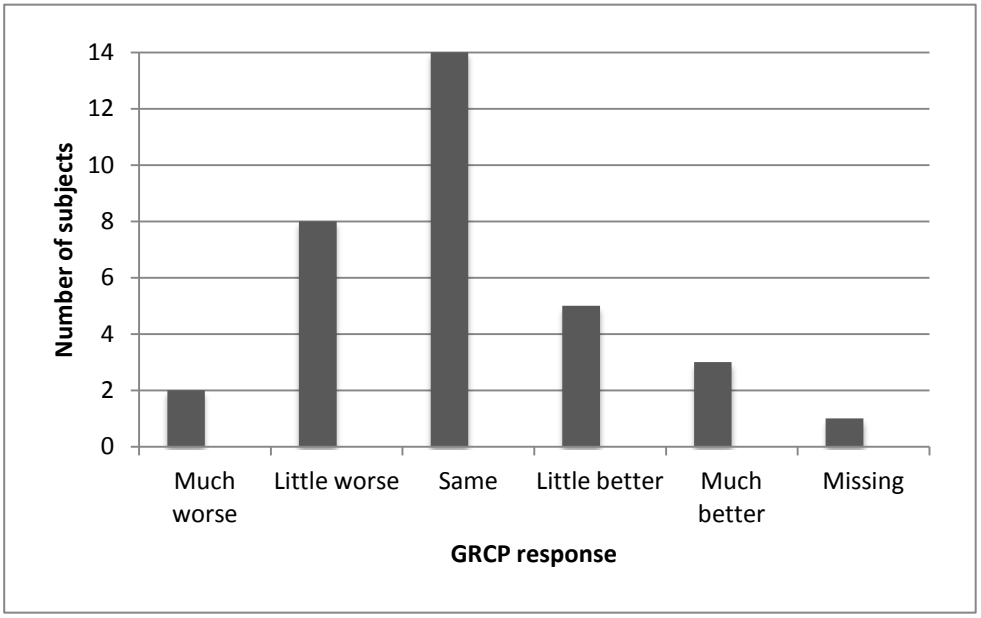
Combining group B (n=21) and the pre-injection data from group $A_{pre\text{-}post}$ (n=12) into the 'not injected' reliability group gave a sample size of 33. There were no missing SUPER-KIDZ data for this analysis, however one patient did not complete the GRCP (inferred as "not the same" in the secondary analysis). The 2- and 3- week post-injection data from Group A gave a sample size of 30 for the 'injected' reliability group, of which 4 patients were missing SUPER-KIDZ data for either or both of the 2- or 3- weeks post- injection time points. Of the remaining 26 patients, 1 did not complete the GRCP (inferred as "not the same" in the secondary analysis).

3.4.2 Distribution of GRCP responses

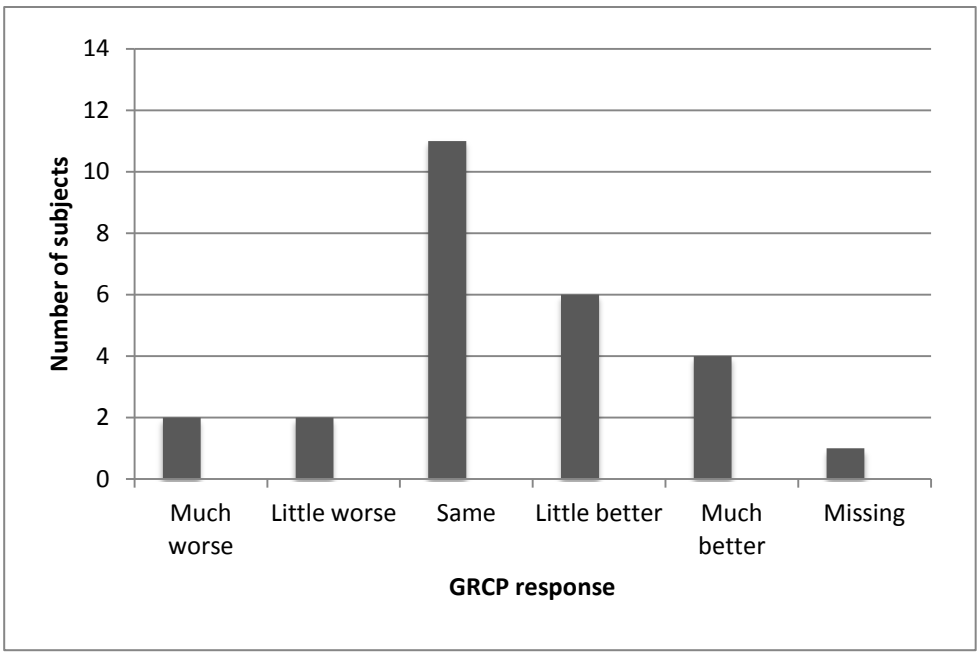
Figure 6 shows the distribution of GRCP responses between the test-retest reliability time points for each reliability group. Of the 33 patients in the 'not injected' subgroup, 14 (42%) reported stable pain. From the 'injected' reliability subgroup 11 of the 26 patients (42%) stated that their pain remained stable between the time points. For each ICC and kappa parameter, the primary analysis was performed in all study subjects regardless of their GRCP response, and a secondary analysis was performed only in those reporting stable pain.

Figure 6: Distribution of GRCP responses for reliability time points in each analysis group

a) In the 'not injected' reliability group (n=33), 14 patients reported stable pain



b) In the ‘injected’ reliability group (total n=26), 11 patients reported stable pain



3.4.3 Continuous variables: Intra-class correlation coefficient

When all subjects were analyzed, the ICC (2,1) values for the ‘not injected’ group were between 0.74 and 0.91, and those for the ‘injected’ group ranged between 0.68-0.86 (Table 12). The ICC for Q2 and Q5 in the ‘not injected’ group, and for Q5 in the ‘injected’ group met the hypothesized criteria in H1 of ≥0.80 and a lower bound of >0.60 in the 95% confidence interval.

When only the stable patients were analyzed, all of the ICCs values increased compared to the primary analysis except for the ICC of Q5 in the ‘not injected’ group (0.70 compared with 0.91). The ICC for Q2 (0.87) in the stable ‘not injected’ group and the ICC for all 3 items (0.81-0.91) in the stable ‘injected’ group achieved the hypothesized criteria of ≥0.80. However, the 95% confidence intervals for Q1 and Q2 in the ‘injected’ group did cross 0.60.

Thus, all 3 continuous items met the criteria in H1 in at least one analysis, and all of the ICC (2,1) values were greater than or equal to 0.68.

Table 12: Intra-class correlation coefficients (ICC) for continuous items (Q1, Q2, Q5) of SUPER-KIDZ questionnaire for the two reliability subgroups

a) 'Not injected' reliability subgroup (all n=33, stable n=14)

Item	Question	GRCP	Mean Test 1	Mean Test 2	ΔScore	ICC (2,1)	95% CI	ICC≥0.80 & lower CI >0.60
Q1	Current pain intensity	All	4.12	3.42	-0.70	0.74	0.53-8.86	N
		Stable	3.79	2.86	-0.93	0.76	0.32-0.92	N
Q2	Average pain over past week	All	4.76	4.27	-0.51	0.82	0.67-0.91	Y
		Stable	4.85	4.50	-0.35	0.87	0.65-0.96	Y
Q5	# painful body locations	All	7.76	8.33	0.57	0.91	0.82-0.95	Y
		Stable	6.79	6.64	-0.15	0.70	0.30-0.90	N

GRCP=5-point global rating of change in pain; ICC=intra-class correlation coefficient; CI=confidence interval

b) 'Injected' reliability subgroup (all n=26, stable n=11)

Item	Question	GRCP	Mean Test 1	Mean Test 2	ΔScore	ICC (2,1)	95% CI	ICC≥0.80 & lower CI >0.60
Q1	Current pain intensity	All	2.50	2.85	0.35	0.68	0.41-0.84	N
		Stable	3.00	2.91	-0.09	0.82	0.45-0.95	N
Q2	Average pain over past week	All	3.12	3.58	0.36	0.70	0.44-0.86	N
		Stable	3.00	3.18	0.18	0.81	0.42-0.94	N
Q5	# painful body locations	All	2.85	4.80	1.95	0.86	0.70-0.93	Y
		Stable	3.81	4.81	1.00	0.91	0.71-0.98	Y

GRCP=5-point global rating of change in pain; ICC=intra-class correlation coefficient; CI=confidence interval

3.4.4 Continuous variables: Error parameters and minimal detectable change

The SEM expressed in the units of the SUPER-KIDZ measure for items Q1, Q2, and Q5 ranged between 1.08 and 2.97 (Table 13). As expected, the largest SEM corresponded to Q5, given the larger variance of the responses to this question. The MDC_{95} for both Q1 and Q2 was a change in score of between 3 and 4 units on the 11-point scale, and the MDC_{95} for Q5 was somewhat higher corresponding to a change score of about 6 to 8 body locations. This indicates that, 95% of the time, a change in item score greater than or equal to these values would be reflective of a true difference, given the measurement error of the test.

Table 13: Summary of error parameters and minimal detectable change for continuous SUPER-KIDZ item scores

Item	Question	Group	ICC (2,1)	SS_{TOTAL}	SEM	SE_{diff}	MDC_{95}
Q1	Current pain intensity	Not injected	0.74	248.62	1.42	2.00	3.92
		Injected	0.68	116.28	1.22	1.72	3.37
Q2	Average pain over past week	Not injected	0.82	207.36	1.08	1.52	2.99
		Injected	0.70	140.83	1.30	1.83	3.59
Q5	# Painful body locations	Not injected	0.91	3136.3	2.97	4.19	8.21
		Injected	0.86	896.00	2.24	3.16	6.19

ICC=intra-class correlation coefficient; SS_{TOTAL}= total sums of squares from corresponding ANOVA; SEM=standard error of measurement, SE_{diff}=standard error of difference scores; MDC_{95}=minimal detectable change at 95% confidence level

3.4.5 Ordinal variables: Weighted Cohen's kappa

When all subjects were analyzed, the weighted Cohen's kappa values tended to be greater in magnitude in the 'injected' group compared with the 'not injected' group. In the former, 9 of the 20 kappas (Q3, Q4, Q8, Q9, Q11-13, Q16, Q17) met the pre-specified criteria in H1 ($\geq$0.80) while in the latter only 2 kappas (Q3, Q4) achieved the hypothesized criteria (Table 14 & Table 15). Also, 5 additional kappas in the 'injected' group were between 0.70-0.79, compared with 1 in the 'not injected' group.

When only the stable patients were analyzed, many of the kappas decreased in magnitude. However, some of the item kappas were similar or improved, namely Q3, Q6, Q11, Q19 and Q20 in the 'injected' group. In this analysis, 2 questions in the 'not injected' group (Q3, Q4) and 8 questions in the 'injected' group (Q3, Q4, Q6, Q11, Q13, Q17, Q19, Q20) met the hypothesized criteria in H1.

The lowest weighted kappas in the primary analysis in the 'not injected' group corresponded to Q14 (κ=0.48), Q17 (κ=0.47), and Q18 (κ=0.41). In the 'injected' group, Q7 had the lowest kappa (κ=0.45), and due to all responses being zero at the first time point, had a kappa of zero in the secondary analysis.

Selected cross-tables are presented in Figure 7 to illustrate specific situations of interest for the weighted kappa analysis.

Table 14: Weighted Cohen's Kappa for ordinal items of SUPER-KIDZ questionnaire in the 'not injected' reliability group (all n=33, stable n=14)

Domain	Item	Question	GRCP	% agree-ment	Wt kappa	95% CI	κ≥0.8 & lower CI >0.6
Pain characteristics	Q3	Pain frequency	All	0.55	0.87	0.79-0.96	Y
			Stable	0.50	0.85	0.71-0.99	Y
	Q4	Pain duration	All	0.61	0.87	0.79-0.96	Y
			Stable	0.57	0.86	0.72-1.00	Y
Assoc'd symptoms	Q6	Fatigue frequency	All	0.48	0.50	0.21-0.80	N
			Stable	0.57	0.65	0.37-0.93	N
Interference	Q7	Difficulty sleeping	All	0.45	0.61	0.37-0.85	N
			Stable	0.43	0.34	-0.06-0.73	N
	Q8	Difficulty paying attention	All	0.45	0.66	0.49-0.83	N
			Stable	0.57	0.49	0.16-0.81	N
	Q9	Difficulty standing	All	0.48	0.71	0.53-0.89	N
			Stable	0.29	0.39	0.03-0.76	N
	Q10	Difficulty having fun	All	0.48	0.68	0.49-0.87	N
			Stable	0.27	0.31	-0.15-0.77	N
	Q11	Difficulty with schoolwork	All	0.45	0.60	0.39-0.81	N
			Stable	0.43	0.38	-0.14-0.91	N
	Q12	Difficulty walking 1 block	All	0.45	0.62	0.35-0.90	N
			Stable	0.36	0.34	-0.18-0.87	N
	Q13	Difficulty running	All	0.52	0.66	0.45-0.86	N
			Stable	0.29	0.33	-0.03-0.70	N
	Q14	How much want pain to stop	All	0.27	0.48	0.18-0.78	N
			Stable	0.14	0.22	-0.31-0.74	N
	Q15	Afraid pain would get worse	All	0.45	0.61	0.43-0.80	N
			Stable	0.21	0.31	0.07-0.55	N
	Q16	Can't stand it anymore	All	0.55	0.64	0.35-0.94	N
			Stable	0.29	0.24	-0.35-0.83	N
Emotional functioning	Q17	Feeling sad	All	0.42	0.47	0.20-0.74	N
			Stable	0.43	0.29	-0.18-0.75	N
	Q18	Feeling angry	All	0.48	0.41	0.11-0.71	N
			Stable	0.50	-0.01	-0.52-0.51	N
	Q19	Feeling cheerful	All	0.42	0.52	0.22-0.81	N
			Stable	0.29	0.52	0.17-0.87	N
	Q20	Feeling worried	All	0.31	0.50	0.22-0.78	N
			Stable	0.29	0.29	-0.29-0.83	N

Wt kappa=quadratic weighted kappa; CI=confidence interval

Table 15: Weighted Cohen's Kappa for ordinal items of SUPER-KIDZ questionnaire from the 'injected' reliability group (all n=26, stable n=11)

Domain	Item	Question	GRCP	% agree-ment	Wt kappa	95% CI	κ≥0.8 & lower CI >0.6
Pain characteristics	Q3	Pain frequency	All	0.56	0.84	0.70-0.98	Y
			Stable	0.82	0.97	0.92-1.00	Y
	Q4	Pain duration	All	0.58	0.83	0.68-0.98	Y
			Stable	0.45	0.86	0.74-0.99	Y
Assoc'd symptoms	Q6	Fatigue frequency	All	0.58	0.66	0.42-0.90	N
			Stable	0.73	0.90	0.75-1.00	Y
Inteference	Q7	Difficulty sleeping	All	0.65	0.45	0.05-0.85	N
			Stable[§]	0.64	0.00	0.00-0.00	N
	Q8	Difficulty paying attention	All	0.73	0.86	0.71-1.00	Y
			Stable	0.73	0.78	0.56-1.00	N
	Q9	Difficulty standing	All	0.62	0.86	0.77-0.96	Y
			Stable	0.55	0.74	0.54-0.94	N
	Q10	Difficulty having fun	All	0.62	0.77	0.59-0.95	N
			Stable	0.50	0.67	0.38-0.95	N
	Q11	Difficulty with schoolwork	All	0.77	0.84	0.70-0.99	Y
			Stable	0.73	0.92	0.85-0.98	Y
	Q12	Difficulty walking 1 block	All	0.65	0.84	0.71-0.96	Y
			Stable	0.64	0.56	0.34-0.78	N
	Q13	Difficulty running	All	0.62	0.85	0.73-0.96	Y
			Stable	0.65	0.90	0.80-1.00	Y
	Q14	How much want pain to stop	All	0.69	0.70	0.45-0.95	N
			Stable	0.82	0.49	-0.05-1.00	N
	Q15	Afraid pain would get worse	All	0.65	0.61	0.28-0.94	N
			Stable	0.64	0.27	-0.26-0.79	N
	Q16	Can't stand it anymore	All	0.81	0.83	0.63-1.00	Y
			Stable	0.82	0.42	-0.17-1.00	N
Emotional functioning	Q17	Feeling sad	All	0.77	0.83	0.66-0.99	Y
			Stable	0.73	0.85	0.66-1.00	Y
	Q18	Feeling angry	All	0.69	0.74	0.50-0.98	N
			Stable	0.64	0.76	0.58-0.93	N
	Q19	Feeling cheerful	All	0.58	0.70	0.46-0.94	N
			Stable	0.73	0.90	0.75-1.00	Y
	Q20	Feeling worried	All	0.62	0.75	0.54-0.97	N
			Stable	0.73	0.88	0.70-1.00	Y

[§]For Q7, all of the 1st time point responses=0; Wt kappa=quadratic weighted kappa; CI=confidence interval

Figure 7: Cross-plots of responses for selected SUPER-KIDZ items at 2 reliability time points

a) Question 14 - stable subjects, 'not injected' group (n=14): low kappa (κ=0.22) due to poor marginal distribution despite heterogeneous response distribution

Q14	Q14 Time 2					
Time 1	0	1	2	3	4	Total
0	0	1	0	0	1	2
1	2	1	0	0	0	3
2	0	0	0	0	1	1
3	0	1	1	1	1	4
4	1	0	0	3	0	4
Total	3	3	1	4	3	14

b) Question 13 - stable subjects, 'injected' group (n=11): heterogeneous response distribution and good marginal distribution gives high kappa (κ=0.90) despite small sample size

Q13	Q13 Time 2					
Time 1	0	1	2	3	4	Total
0	2	1	0	0	0	3
1	1	2	0	0	0	3
2	0	0	0	1	0	1
3	0	0	1	2	0	3
4	0	0	0	0	1	1
Total	3	3	1	3	3	11

c) Question 16 - all subjects, 'injected' group (n=26): relatively homogeneous population (mostly '0' and '1') but reasonable marginal distribution gives high kappa (κ=0.83)

Q16	Q16 Time 2					
Time 1	0	1	2	3	4	Total
0	13	3	0	0	0	16
1	0	6	1	1	0	8
2	0	0	1	0	0	1
3	0	0	0	0	0	0
4	0	0	0	0	1	1
Total	13	9	2	1	1	26

d) Question 16 - stable subjects, 'injected' group (n=11): very homogeneous population (all responses '0' or '1') with poor marginal distribution results in lower kappa (κ=0.42)

Q16	Q16 Time 2					
Time 1	0	1	2	3	4	Total
0	8	2	0	0	0	10
1	0	1	0	0	0	1
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
Total	8	3	0	0	0	11

**Note that both examples c) and d) have the same percent agreement (0.81 and 0.82)

3.4.6 Summary of test-rest reliability analyses

For the primary analysis, 11 SUPER-KIDZ items achieved the H1 hypothesis of a reliability coefficient of ≥0.80 in either or both of the reliability subgroups ('not injected' or 'injected'). Seven of these items also met the criteria in the secondary analysis of self-reported stable patients, along with an additional 4 items. Thus, 15 SUPER-KIDZ items met the test-retest reliability criteria specified in H1 in at least one analysis. The sensory items Q2-Q5 were the most consistently reliable items (Table 16).

Table 16: Summary of test-retest reliability results for SUPER-KIDZ items in the primary and secondary analyses

Domain	**Item**	**Question**	**Primary analysis (all patients)**	**Secondary analysis (stable patients)**
Pain characteristics	Q1	Current pain	No	Yes
	Q2	Avg pain past week	Yes	Yes
	Q3	Pain frequency	Yes	Yes
	Q4	Pain duration	Yes	Yes
	Q5	# Painful locations	Yes	Yes
Associated	Q6	Fatigue frequency	No	Yes
Interference	Q7*	Sleeping	No	No
	Q8	Paying attention	Yes	No
	Q9	Standing	Yes	No
	Q10*	Having fun	No	No
	Q11	Schoolwork	Yes	Yes
	Q12	Walk 1 block	Yes	No
	Q13	Running	Yes	Yes
	Q14*	Want pain to stop	No	No
	Q15*	Afraid get worse	No	No
	Q16	Can't stand it anymore	Yes	No
Emotional functioning	Q17	Feel sad	Yes	Yes
	Q18*	Feel angry	No	No
	Q19	Feel cheerful	No	Yes
	Q20	Feel worried	No	Yes
Totals:			**Yes=11 items**	**Yes=11 items**
			Either primary or secondary: Yes =15 items **Both primary & secondary: Yes=7 items**	

Yes = ICC or κ ≥0.80 in either or both reliability groups, No = ICC or κ <0.80 in both subgroups; *Items did not meet criteria for test-retest reliability in any analysis

3.5 Responsiveness

3.5.1 Missing data and final sample size for responsiveness analyses

The total sample size for Group A was 30. One patient (subject #43) experienced a complication in the first week post-joint injection (steroid crystallization) causing significantly increased pain, but was better by 2 weeks post-injection. Thus, this participant's 1-week post-injection time point was excluded from the responsiveness data, leaving 29 subjects for analysis at 1 week. All of the subjects completed the required SUPER-KIDZ questionnaires for the 1-week responsiveness analysis, however the GRCP was missing for one of these subjects (inferred as "not improved" in the secondary analysis). At 2 weeks post-injection 3 subjects missed completing the SUPER-KIDZ questionnaire, leaving 27 patients for this analysis. Of these, no GRCP responses were missing.

For the regression analysis, of 90 possible observations (30 subjects with 3 time points each), 3 were missing and 1 was excluded (subject #43, 1-week post-injection), therefore 86 sets of SUPER-KIDZ questionnaire responses were analyzed.

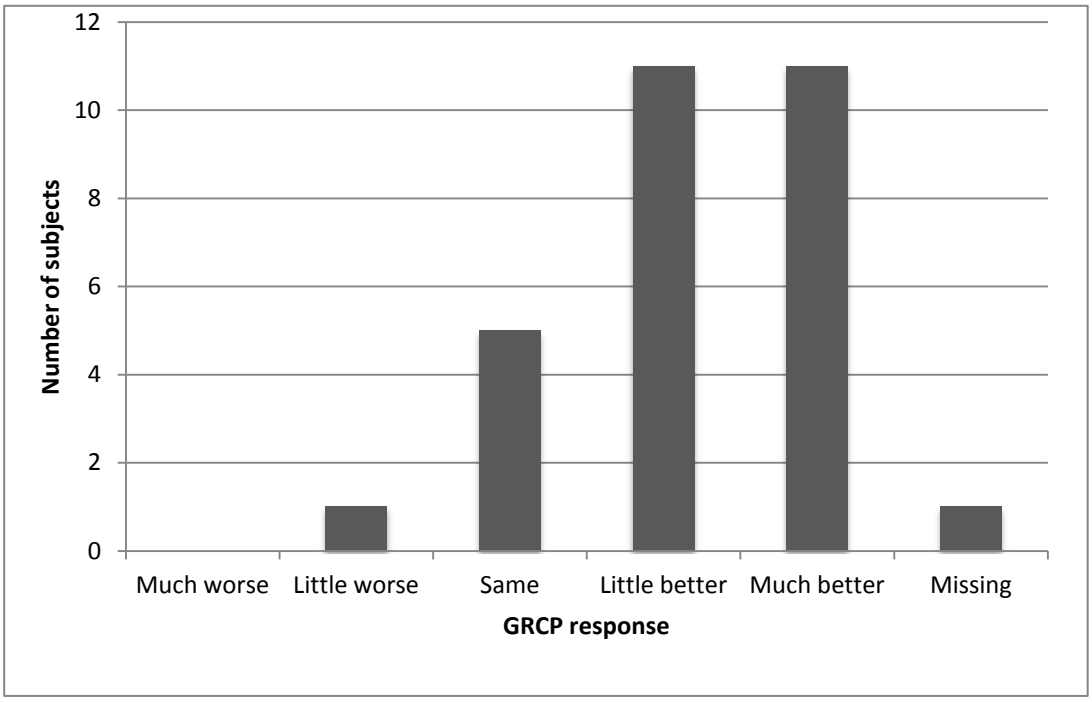
3.5.2 Distribution of GRCP responses

At 1 week after injection, 22/29 (76%) of study subjects reported improvement, defined as "a little better" or "much better". At 2 weeks post-injection, the proportion of patients reporting further improvement ("a little better" or "much better") was 12/27 (44%) (Figure 8). Recalling the criteria for improvement defined in Table 8 (section 2.4.2.2), a total of 22/27 (81%) study subjects met the definition of improvement at 2 weeks post-joint injection (Table 17).

For each time point, the primary responsiveness analysis was performed in all patients undergoing joint injection and a secondary analysis included only those patients who self-reported improvement based on the GRCP responses.

Figure 8: Distribution of GRCP responses for responsiveness time points

a. At 1 week post-injection (n=29), 22 patients reported improvement



b. At 2 weeks post-injection (n=27), 12 patients reported further improvement

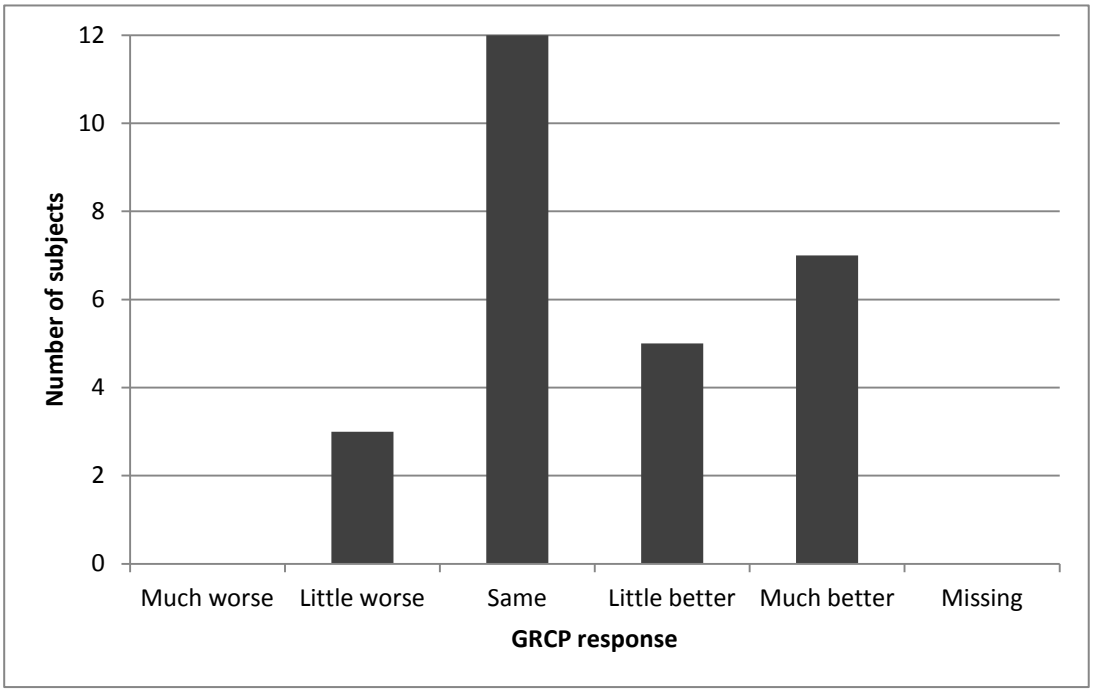


Table 17: Global Rating of Change in Pain (GRCP) responses at 1 and 2 weeks after joint injection (JI)

GRCP* at 1 week after JI	**GRCP* at 2 weeks after JI**	**Meet criteria for improvement at 2 weeks?**	**Number of patients (total n=27)**
Any worse	Any worse	No	0
Any worse	Same	No	0
A little worse	A little better	No	1
A little worse	Much better	Yes	0
Much worse	A little better	No	1
Much worse	Much better	No	0
Same	Any worse	No	0
Same	Same	No	2
Same	Any better	Yes	4
A little better	A little worse	No	1
A little better	Much worse	No	0
Much better	A little worse	Yes	2
Much better	Much worse	No	0
Any better	Same	Yes	10
Any better	Any better	Yes	6
IMPROVED AT 2 WEEKS POST-JI:		Total Yes	22
		Total No	5
		Total missing	0

*GRCP is answered relative to the previous 1 week time period
NB. Any better = much or a little better; Any worse = much or a little worse

3.5.3 Distribution of SUPER-KIDZ item scores over time

Boxplots for each of the SUPER-KIDZ items were constructed to see the trend of scores over the relevant time points for the responsiveness analysis: day before joint injection (‘visit 2’), 1-week post-injection (‘visit 3’), and 2-weeks post-injection (‘visit 4’) (Appendix 8). Based on the graphs, the general trend appears to be a reduction in scores for each item at visits 3 and/or 4 compared with visit 2, except for Q6 and Q19, whose scores appear to stay relatively constant after the joint injection despite being mid-range at baseline. As described in section 3.2, ceiling effects are noted in the baseline score distributions of Q3, Q4, Q13, Q14, and more mildly for Q16. Not surprisingly, these items appear to improve after the joint injection. Low baseline responses (floor effects) are seen for Q7 and Q8, however these distributions still show a trend for improvement in scores after injection. Questions #11, 12, 17, 18, and 20 also have low baseline scores and show a milder graphical trend towards improvement.

The change scores at 2-weeks post-joint injection, compared to the day before injection was larger in the subjects who reported improvement in pain compared with those who did not improve for most of the SUPER-KIDZ items (Table and graphical distributions in Appendix 8). The exceptions include Q6 (fatigue frequency), Q17 (feeling sad), and Q19 (feeling cheerful), for which the unimproved individuals had higher change scores. Of course, this finding needs to be interpreted in light of the very small sample size not reporting improvement (n=5). Items with the largest mean change scores (>0.90) included the sensory items Q1-Q5, and interference items Q12-Q16. The smallest mean change scores (<0.30) corresponded to items Q6 (fatigue), Q17-Q20 (emotional items). The overall range of change scores overlapped between the two groups for all of the SUPER-KIDZ items. The IQRs for the change scores were most distinct between the improved and not improved groups for Q1-Q4, Q12 and Q13; the IQRs overlapped for the remaining items.

3.5.4 Continuous variables: Standardized response mean

The SRM for the continuous items were calculated from the mean difference and SD of the difference according to Equation 12 (Table 18). As hypothesized in H2a, the SRMs are low-to-moderate (0.33-0.50) at 1 week, and are moderate-to-high at 2 weeks post-injection (0.66-0.82). All three items met the criteria in H2a. Question 5 was most responsive at 1 week (SRM=0.50) and question 1 was most responsive at 2 weeks (SRM=0.82).

When only those patients reporting improvement based on their GRCP are analyzed, the magnitude of the SRM increased for Q1 and Q2 at both time points, as expected. However the SRM for Q5 remained about the same in this group based on similar mean differences and relatively larger SD.

Table 18: Mean differences in scores, standard deviation of differences, and standardized response means for continuous SUPER-KIDZ items

a) at 1-week post-joint injection (all n=29, improved n=22)

Item	Question	GRCP	Mean Pre-JI	Mean 1 week post-JI	Mean diff	SD diff	SRM	95% CI
Q1	Current pain intensity	All	4.07	2.86	-1.21	2.65	0.46	0.10-0.72
		Improved	4.50	2.59	-1.91	2.27	0.84	0.42-1.26
Q2	Average pain intensity over past week	All	4.45	3.69	-0.76	2.31	0.33	-0.03-0.69
		Improved	4.63	3.36	-1.27	2.39	0.53	0.11-0.95
Q5	Number of painful body locations	All	6.34	4.62	-1.72	3.47	0.50	0.14-0.86
		Improved	6.36	4.32	-2.05	3.76	0.55	0.13-0.97

SRM=standardized response mean; JI=joint injection; SD=standard deviation; CI=confidence interval

b) at 2 weeks post-joint injection (all n=27, improved n=22)

Item	Question	GRCP	Mean Pre-JI	Mean 2 weeks post-JI	Mean diff	SD diff	SRM	95% CI
Q1	Current pain intensity	All	4.22	2.41	-1.81	2.20	0.82	0.44-1.20
		Improved	4.59	2.32	-2.27	2.16	1.05	0.62-1.48
Q2	Average pain intensity over past week	All	4.70	3.00	-1.70	2.45	0.69	0.31-1.07
		Improved	4.91	2.86	-2.05	2.26	0.91	0.48-1.34
Q5	Number of painful body locations	All	6.86	4.56	-2.30	3.50	0.66	0.28-1.04
		Improved	6.45	4.00	-2.45	3.84	0.64	0.21-1.07

SRM=standardized response mean; JI=joint injection; SD=standard deviation; CI=confidence interval

3.5.5 All variables: Wilcoxon signed rank analysis

When the change in scores of both the continuous and ordinal items were analyzed using the Wilcoxon signed rank test, 7/20 items (Q1, Q5, Q8, Q13-Q16) showed statistically significant change at 1 week post-injection, and 16/20 (80%) items (Q1-Q5, Q7-Q17) changed in a statistically significant way at 2 weeks, which meets the criteria specified in H2b. Seven items had p-values of 0.001 or smaller at 2 weeks (Table 19). The least responsive items included Q6 (fatigue), and 3 of the emotional items (Q18: angry, Q19: cheerful and Q20: worried).

Table 19: Results of Wilcoxon signed rank tests at 1 and 2 weeks post-joint injection (JI)

Item	**Question**	**1 week post-JI (n=29) (expect mild improvement)**			**2 weeks post-JI (n=27) (expect moderate-large improvement)**		
		Median diff (IQR)	S value	p-value	Median diff (IQR)	S value	p-value
Q1	Current pain	1.0 (0.0-3.0)	73.0	0.02	1.0 (0.0-4.0)	104.0	<.0001
Q2	Average pain	0.0 (-1.0-3.0)	44.5	0.12	2.0 (0.0-3.0)	94.0	0.0007
Q3	Pain frequency	0.0 (0.0-1.0)	25.5	0.15	0.0 (0.0-2.0)	41.5	0.05
Q4	Pain duration	0.0 (0.0-0.0)	8.0	0.56	1.0 (0.0-2.0)	71.0	0.003
Q5	#Painful locations	0.5 (-1.0-4.0)	79.5	0.006	2.0 (0.0-4.0)	104.0	<.0001
Q6	Fatigue frequency	0.0 (0.0-1.0)	25.5	0.16	0.0 (0.0-1.0)	13.0	0.38
Q7	Sleeping	0.0 (0.0-0.0)	11.5	0.37	0.0 (0.0-1.0)	27.0	0.008
Q8	Paying attention	0.0 (0.0-1.0)	30.0	0.02	0.0 (0.0-1.0)	47.0	0.01
Q9	Standing	0.0 (-1.0-1.0)	38.0	0.21	0.0 (0.0-2.0)	42.0	0.006
Q10	Having fun	0.0 (0.0-1.0)	45.5	0.06	0.0 (0.0-1.0)	56.0	0.002
Q11	Schoolwork	0.0 (0.0-0.0)	7.5	0.64	0.0 (0.0-1.0)	27.0	0.03
Q12	Walk 1 block	0.0 (0.0-1.0)	18.5	0.31	1.0 (0.0-2.0)	52.5	0.02
Q13	Running	0.5 (-1.0-2.0)	74.0	0.02	1.0 (0.0-2.0)	82.0	0.001
Q14	Want pain to stop	1.0 (0.0-2.0)	75.5	0.003	2.0 (0.0-3.0)	76.5	<.0001
Q15	Afraid pain worse	0.0 (0.0-1.0)	54.5	0.007	1.0 (0.0-2.0)	69.0	0.001
Q16	Can't stand it	0.0 (0.0-1.0)	38.5	0.03	0.0 (0.0-1.0)	70.5	0.0002
Q17	Feel sad	0.0 (0.0-1.0)	20.0	0.30	0.0 (0.0-1.0)	28.0	0.03
Q18	Feel angry	0.0 (0.0-0.0)	11.5	0.24	0.0 (0.0-0.0)	9.0	0.40
Q19	Feel cheerful	0.0 (-1.0-0.0)	-5.5	0.86	0.0 (0.0-1.0)	11.5	0.45
Q20	Feel worried	0.0 (0.0-1.0)	27.0	0.22	0.0 (0.0-1.0)	26.0	0.08

IQR=interquartile range

3.5.6 Regression analysis

3.5.6.1 Initial linear mixed model regression

Based on smallest AIC parameter, most of the linear mixed models fit best with a compound symmetry variance matrix, however the models for Q5, Q10 and Q16 fit best with an unstructured correlation matrix, Q7 and Q11 fit best with a heterogeneous compound symmetry structure, and question 12 fit best with an autoregressive structure. This implies that the response variation for the majority of items is homogeneous over time, however certain items have a heterogeneous variance that was accordingly specified in the mixed model.

In the original models, the overall effect of time over the whole 2-week period was significant in 13 of the 20 items (Q1-Q5, Q7-Q10, Q13-Q16) (Table 20). At 1 week, scores of 9/20 items significantly improved (Q1, Q2, Q5, Q8, Q10, and Q13-Q16), and at 2 weeks, 16/20 (80%) item scores significantly improved. These findings meet the criteria specified in H4. Half of the p-values were 0.001 or smaller at 2 weeks. Question 6 (fatigue) and emotional items 18-20 (angry, cheerful, worried) were not found to significantly change by 2 weeks. These are the same items that did not show responsiveness when analyzed using the Wilcoxon signed rank test in section 3.5.5 above.

In terms of the magnitude of change in the units of the SUPER-KIDZ measure, the beta coefficients for the items that significantly changed at 2 weeks ranged between 1.7 and 2.2 units for the continuous items (Q1, Q1, Q5) and between 0.36 and 1.48 units for the ordinal variables.

Table 20: Linear mixed models for repeated measures of SUPER-KIDZ items at 1- and 2- weeks post-joint injection (JI) (#observations=86)

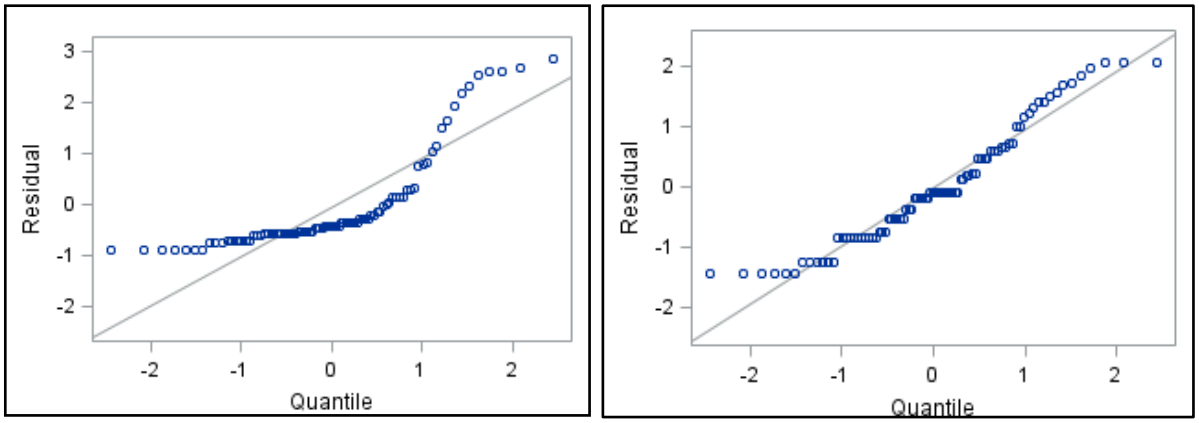
Item	**Question**	**Overall effect of time**		**1 week post-JI (expect mild improvement)***				**2 weeks post-JI (expect moderate-large improvement)***			
		F	p-value	β	SE	95% CI	p-value	β	SE	95% CI	p-value
Q1	Current pain	8.18	0.0008	1.18	0.43	0.32-2.03	0.008	1.72	0.44	0.84-2.60	0.0003
Q2	Average pain	7.90	0.001	0.84	0.41	0.16-1.67	0.05	1.68	0.42	0.83-2.53	0.0002
Q3	Pain frequency	3.12	0.05	0.58	0.33	-0.07-1.25	0.08	0.81	0.34	0.13-1.49	0.02
Q4	Pain duration	6.67	0.003	0.11	0.28	-0.45-0.67	0.70	0.97	0.29	0.40-1.54	0.001
Q5	#Painful locations	6.05	0.006	1.83	0.64	0.52-3.14	0.008	2.19	0.64	0.89-3.49	0.002
Q6	Fatigue frequency	1.31	0.28	0.29	0.18	-0.07-0.66	0.11	0.14	0.19	-0.23-0.52	0.44
Q7	Sleeping	5.95	0.005	0.15	0.15	-0.16-0.43	0.37	0.47	0.14	0.19-0.76	0.002
Q8	Paying attention	4.46	0.02	0.46	0.18	0.11-0.82	0.01	0.45	0.18	0.09-0.82	0.02
Q9	Standing	3.53	0.04	0.38	0.25	-0.12-0.89	0.14	0.68	0.26	0.16-1.20	0.01
Q10	Having fun	6.86	0.004	0.53	0.26	0.01-1.06	0.05	0.69	0.19	0.31-1.08	0.001
Q11	Schoolwork	2.88	0.06	0.13	0.17	-0.21-0.48	0.44	0.39	0.16	0.06-0.71	0.02
Q12	Walk 1 block	2.79	0.07	0.20	0.24	-0.28-0.67	0.40	0.68	0.30	0.08-1.28	0.03
Q13	Running	6.94	0.002	0.64	0.25	0.13-1.15	0.01	0.94	0.26	0.42-1.46	0.0007
Q14	Want pain to stop	16.8	<.0001	0.97	0.26	0.46-1.48	0.0004	1.48	0.26	0.96-2.00	<.0001
Q15	Afraid pain worse	9.02	0.0004	0.55	0.20	0.16-0.95	0.007	0.83	0.20	0.43-1.24	0.0001
Q16	Can't stand it	12.0	0.0002	0.59	0.22	0.13-1.04	0.01	0.88	0.20	0.47-1.28	0.0001
Q17	Feel sad	2.76	0.07	0.21	0.15	-0.09-0.51	0.17	0.36	0.15	0.05-0.66	0.02
Q18	Feel angry	1.01	0.37	0.17	0.13	-0.09-0.42	0.19	0.14	0.13	-0.12-0.40	0.29
Q19	Feel cheerful	0.52	0.60	-0.06	0.21	-0.49-0.37	0.77	0.16	0.22	-0.28-0.60	0.47
Q20	Feel worried	2.04	0.14	0.29	0.16	-0.04-0.61	0.08	0.28	0.16	-0.05-0.61	0.10

β =beta coefficient from regression model; SE=standard error; CI=confidence interval; *as specified in H4

3.5.6.2 Testing assumptions: normality of residuals and homoscedasticity

Examining the quantile-quantile plots (Appendix 9) for each model, the residuals appear normally distributed for Q1, Q2 and Q14. Questions 4, 6 and 9 also have reasonably straight residual plots. The remaining items have straight left tails due to the relatively large number of zero responses on the ordinal scale. Some also plateau at the larger values. Unfortunately logarithmic transformation was not possible for these models due to the zero values. Square root and exponential transformations were performed and resulted in no change to the quantile-quantile plots since the many zero values still had the same value after transformation. In the case of Q5, there was only 1 zero response. As such, a log-transformation was possible (zero value removed), and resulted in an improved quantile-quantile plot (Figure 9). In addition the AIC decreased from 473.27 to 176.24. The disadvantage is that the model parameters need to be exponentiated (e^x) in order to interpret the results.

Figure 9: Quantile-quantile plot of residuals for question 5 model before (left) and after (right) log-transformation of dependent variable



Homoscedasticity of the residuals was difficult to examine because of only 3 distinct time points and the ordinal nature of the data. In general, the plots of studentized residuals versus predicted values were symmetrically distributed among the 3 points and none of the models had gross heteroscedasticity (Appendix 9).

3.5.6.3 Identification of potentially influential outliers

Several potentially influential observations were identified based on Cook's D criteria (> 4/n = 0.05 for n of 86). A priori, it was decided to look for observations that significantly changed model parameters (i.e. *F* statistic, β coefficient), and to then assess whether they should be

removed based on several factors including the magnitude of the difference between predicted and observed values, the clinical information given by the subject, and whether the observations led to more or less conservative parameter estimates. The models for Q2, Q3, Q9, Q12, Q17 and Q20 contained observations that significantly changed the model parameters upon removal (Table 21), and these specific changes are highlighted in Table 22.

Table 21: Potentially influential observations for linear mixed model regression and impact on the model parameters

Item	Subject (visit #)*	Obs value	Pred value	Cook's D	Change in model parameters upon removal of influential observation
1- Curr pain	27 (visit 4)	7	2.3	0.09	No change
2- Avg pain	36 (visit 2)	8	2.9	0.06	**1 week β insignificant**
3- Pain frequency	26 (visit 2)	6	4.6	0.08	**Overall model insignificant**
	36 (visit 2)	7	4.6	0.07	**Overall model & 2 week β insignificant**
4- Duration	31 (visit 4)	1	3.4	0.11	No change
5[§]- Locations	48 (visit 4)	0	0.9	0.14	No change
6- Fatigue frequency	28 (visit 4)	4	2.1	0.08	No change
	28 (visit 3)	1	2.0	0.08	No change
7- Sleeping	33 (visit 4)	3	0.4	0.10	No change
8- Pay attn	38 (visit 3)	4	0.7	0.09	No change
9- Standing	26 (visit 2)	4	1.7	0.09	**Overall model insignificant**
	43 (visit 4)	4	1.0	0.06	**No change**
10- Have fun	31 (visit 3)	3	1.2	0.08	No change
	31 (visit 4)	0	1.0	0.08	No change
	43 (visit 4)	4	1.0	0.08	No change
11- School	33 (visit 2)	4	1.1	0.10	No change
12- Walk 1 block	43 (visit 4)	4	0.9	0.10	**Overall model significant**
	46 (visit 2)	4	1.6	0.09	No change
13- Running	26 (visit 2)	3	1.4	0.07	No change
	27 (visit 4)	4	2.4	0.07	No change
14- Think	29 (visit 3)	4	1.4	0.07	No change
15- Worse	43 (visit 4)	4	0.8	0.13	No change
16- Can't stand it	29 (visit 3)	2	0.8	0.10	No change
	29 (visit 4)	0	0.6	0.10	No change
	40 (visit 2)	4	1.4	0.10	No change
17- Sad	29 (visit 2)	0	1.2	0.08	**Overall model significant**
	24 (visit 3)	1	1.0	0.07	**Overall model significant**
18- Angry	25 (visit 2)	3	1.1	0.12	No change
19- Cheerful	31 (visit 4)	4	1.4	0.10	No change
20- Worried	30 (visit 3)	2	1.1	0.09	**Overall model & 1 week β significant**

*If >1 influential observation, each removed sequentially starting with largest Cook's D. If no change noted with first observation, then stop; [§]Log-transformed model for Q5

Table 22: Key changes in linear mixed model parameters (boxed values) upon removal of influential observations

Item	Overall effect of time		1 week post-JI				2 weeks post-JI			
	F	p-value	β	SE	95% CI	p-value	β	SE	95% CI	p-value
2- avg pain	7.00	0.002	0.70	0.40	-0.10-1.51	0.08	1.54	0.41	0.71-2.36	0.0005
3*- freq	1.57	0.22	0.31	0.29	-0.28-0.90	0.30	0.53	0.30	-0.07-1.14	0.08
9§-standing	2.80	0.07	0.27	0.23	-0.20-0.74	0.26	0.57	0.24	0.09-1.05	0.02
9*-standing	3.74	0.03	0.30	0.23	-0.17-0.76	0.21	0.66	0.24	0.18-1.14	0.009
12- walk	4.30	0.02	0.25	0.22	-0.19-0.70	0.26	0.81	0.29	0.24-1.39	0.006
17*-sad	4.90	0.01	0.20	0.13	-0.06-0.47	0.13	0.42	0.13	0.15-0.69	0.002
20-worried	3.08	0.05	0.35	0.15	0.05-0.66	0.02	0.27	0.15	-0.03-0.58	0.08

JI=joint injection; β =beta coefficient from regression model; SE=standard error; §one influential observation (26 [visit 2]) removed; *both influential observations removed

For Q2 and Q3, the influential observations identified were all from visit 2 (one day pre-injection). These observations were at the high end of the response scale, and when they were removed, the overall model and/or β parameters became insignificant. Thus, since the models were so sensitive to these 1 or 2 observations, they were removed from the final models in order to be conservative. For Q9, removal of the high value from subject 26 (visit 2) resulted in an insignificant *F* statistic. However, removal of another influential observation from 2 weeks post-injection (subject #43, visit 4) resulted in no overall change in the model parameters. Thus, the original model was kept for Q9. Subject #43 (visit 4) also reported a high value for Q12, and removal of this data point resulted in the overall model becoming significant (F=2.79, p=0.07 → F=4.30, p=0.02). In this case, we recall that subject #43 was the patient who had a complication at 1 week post-injection and therefore he/she may still have inflated scores at the 2-week time point (visit 4). Thus, this observation was removed for the final Q12 model. For Q17 and Q20, removal of the influential observations resulted in a significant effect of time for the overall models. The observed values were not particularly different form the predicted values, and there was no clinical reason to exclude these observations. Also given that the 1- and 2- week beta coefficients changed very minimally, the original models were retained for Q17 and Q20.

3.5.6.4 Final linear mixed models

In the final regression models, which reflected the changes made after model diagnostics, the overall effect of time over the whole 2-week period was significant in 13 of the 20 items (Q1, Q2, Q4, Q5, Q7-10, Q12-16) (Table 23). At 1 week post-injection, scores of 8/20 items significantly improved (Q1, Q5, Q8, Q10, and Q13-16). At 2 weeks post-injection, 15/20 (75%) item scores significantly improved (Q1, Q2, Q4, Q5, Q7-17), which meets the criteria specified in H4. Nine of the p-values were 0.001 or smaller at 2 weeks. Question #3 (pain frequency), question #6 (fatigue) and emotional items #18-20 (angry, cheerful, worried) were not found to significantly change after injection. Note that Q3 is no longer found to be responsive by the regression method after removal of the influential outliers identified in section 3.5.6.5. This is consistent with the borderline p=0.05 by the Wilcoxon signed rank test in section 3.5.5 above.

The magnitude of change in units of the SUPER-KIDZ measure at 2 weeks ranged between 1.5 and 1.8 units per unit time for the continuous items (Q1, Q2, Q5), and between 0.36 and 1.48 units per unit time for the ordinal variables. The ordinal items with beta coefficients close to a 1-unit decrease in score at 2 weeks included Q4 (pain duration), Q12 (walk 1 block), Q13 (running), Q14 (thinking about how much wanted pain to stop), Q15 (afraid pain would get worse), and Q16 (couldn't stand it anymore). The continuous item with the greatest magnitude of change was Q5 (number of painful body areas) and the most responsive ordinal item was Q14 (β=1.76 and β=1.48 respectively).

Table 23: Final linear mixed models for repeated measures of SUPER-KIDZ items at 1- and 2- weeks post-joint injection (JI) after model diagnostics performed

Item	Question	Overall effect of time		1 week post-JI (expect mild improvement)*				2 weeks post-JI (expect moderate-large improvement)*			
		F	p-value	β	SE	95% CI	p-value	β	SE	95% CI	p-value
Q1	Current pain	8.18	0.0008	1.18	0.43	0.32-2.03	0.008	1.72	0.44	0.84-2.60	0.0003
Q2¶	Average pain	7.00	0.002	0.70	0.40	-0.10-1.51	0.08	1.54	0.41	0.71-2.36	0.0005
Q3¶	Pain frequency	1.57	0.22	0.31	0.29	-0.28-0.90	0.30	0.53	0.30	-0.07-1.14	0.08
Q4	Pain duration	6.67	0.003	0.11	0.28	-0.45-0.67	0.70	0.97	0.29	0.40-1.54	0.001
Q5§	#Painful locations	12.7	0.001	1.34	0.13	1.02-1.74	0.03	1.76	0.12	1.39-2.25	<.0001
Q6	Fatigue frequency	1.31	0.28	0.29	0.18	-0.07-0.66	0.11	0.14	0.19	-0.23-0.52	0.44
Q7	Sleeping	5.95	0.005	0.15	0.15	-0.16-0.43	0.37	0.47	0.14	0.19-0.76	0.002
Q8	Paying attention	4.46	0.02	0.46	0.18	0.11-0.82	0.01	0.45	0.18	0.09-0.82	0.02
Q9	Standing	3.53	0.04	0.38	0.25	-0.12-0.89	0.14	0.68	0.26	0.16-1.20	0.01
Q10	Having fun	6.86	0.004	0.53	0.26	0.01-1.06	0.05	0.69	0.19	0.31-1.08	0.001
Q11	Schoolwork	2.88	0.06	0.13	0.17	-0.21-0.48	0.44	0.39	0.16	0.06-0.71	0.02
Q12¶	Walk 1 block	4.30	0.02	0.25	0.22	-0.19-0.70	0.26	0.81	0.29	0.24-1.39	0.006
Q13	Running	6.94	0.002	0.64	0.25	0.13-1.15	0.01	0.94	0.26	0.42-1.46	0.0007
Q14	Want pain to stop	16.8	<.0001	0.97	0.26	0.46-1.48	0.0004	1.48	0.26	0.96-2.00	<.0001
Q15	Afraid pain worse	9.02	0.0004	0.55	0.20	0.16-0.95	0.007	0.83	0.20	0.43-1.24	0.0001
Q16	Can't stand it	12.0	0.0002	0.59	0.22	0.13-1.04	0.01	0.88	0.20	0.47-1.28	0.0001
Q17	Feel sad	2.76	0.07	0.21	0.15	-0.09-0.51	0.17	0.36	0.15	0.05-0.66	0.02
Q18	Feel angry	1.01	0.37	0.17	0.13	-0.09-0.42	0.19	0.14	0.13	-0.12-0.40	0.29
Q19	Feel cheerful	0.52	0.60	-0.06	0.21	-0.49-0.37	0.77	0.16	0.22	-0.28-0.60	0.47
Q20	Feel worried	2.04	0.14	0.29	0.16	-0.04-0.61	0.08	0.28	0.16	-0.05-0.61	0.10

β =beta coefficient from regression model; SE=standard error; *as specified in H4; ¶influential observations removed;
§log-transformed model (exponentiated β reported)

3.5.7 ROC Curve Analysis

The ROC curve analysis was performed using the 2-week post-injection data since it was more strongly responsive based on the previous responsiveness analyses. The group was divided according to whether subjects had improved (n=22) or not (n=5) according to the external criterion GRCP response. Based on the distribution of change scores (presented in section 3.5.2), it would be expected that Q1-Q4, Q12 and Q13 would be among the strongest items in distinguishing the improved from the not improved patients.

Nine of the 20 items (Q1-4, Q10, Q12-15) met the criteria of an AUC ≥0.70, however the AUC for Q2 and Q15 had a 95% confidence interval including 0.50 (Table 24). Thus, questions about current pain intensity (Q1), pain frequency (Q3), pain duration (Q4), difficulty having fun (Q10), difficulty walking (Q12), difficulty running (Q13) and thinking about how much wanted pain to stop (Q14) were found to be the most responsive items by this analysis.

The corresponding ROC curves are shown in Appendix 10. Three of the models (Q6, Q17, Q19) have an AUC of <0.50, although the confidence intervals all cross 0.50. An AUC of 0.50 would be expected by random chance, indicating that the items are not helpful in distinguishing between patients with improved pain or not, and are unlikely associated with this outcome. An AUC of <0.50 should not occur unless the test is consistently being interpreted incorrectly by the subjects or the item is inversely associated with the external criterion [136]. Interestingly, the least responsive items once again included Q6 and Q19, consistent with the previous responsiveness analyses.

In terms of change scores, the general trend was that while a change of +1 was reasonably sensitive (>0.70) for improvement in some of the items, the sensitivity of this cut-off was fairly poor overall (0.23-0.68). A change score of 0 had much better sensitivity (>0.90) but lower specificity. Accuracy also tended to be highest (0.74-0.85) with a cut-off of 0. This is likely because some of the subjects reporting improvement at the 2-week time point via the GRCP gave the same or worse SUPER-KIDZ scores.

Table 24: Sensitivity, specificity, and accuracy at different levels of change for each SUPER-KIDZ item at 2 weeks post-joint injection (n=27: of which 22 improved and 5 not improved)

Item	Change score (x)	Sensitivity	Specificity	Accuracy	AUC	SE	95% CI	Lower 95% CI ≥0.50
Q1 – Current pain intensity	-1	0.95	0.00	0.78	0.88	0.07	0.74-1.00	YES
	0	0.95	0.40	0.85				
	1	0.82	0.80	0.81				
	2	0.59	1.00	0.67				
	3	0.45	1.00	0.56				
Q2 – Average pain intensity over past week	-1	0.95	0.20	0.81	0.72	0.16	0.40-1.00	NO
	0	0.86	0.40	0.78				
	1	0.73	0.80	0.74				
	2	0.64	0.80	0.67				
	3	0.36	0.80	0.44				
Q3 – Pain frequency	-1	0.91	0.40	0.81	0.86	0.07	0.72-1.00	YES
	0	0.91	0.60	0.85				
	1	0.55	1.00	0.63				
	2	0.45	1.00	0.56				
Q4 – Pain duration	-1	1.00	0.00	0.81	0.79	0.10	0.60-0.99	YES
	0	0.91	0.40	0.81				
	1	0.64	0.80	0.67				
	2	0.36	1.00	0.48				
Q5 – Number of painful body locations	-1	0.95	0.00	0.78	0.57	0.12	0.34-0.80	NO
	0	0.91	0.00	0.74				
	1	0.73	0.20	0.63				
	2	0.64	0.40	0.59				
	3	0.36	0.80	0.44				
	4	0.32	1.00	0.44				
Q6 – Fatigue frequency	-1	0.95	0.00	0.78	0.40	0.11	0.18-0.62	NO
	0	0.82	0.00	0.67				
	1	0.27	0.60	0.33				
	2	0.09	1.00	0.26				
Q7 – Difficulty sleeping	-1	1.00	0.00	0.81	0.60	0.10	0.41-0.80	NO
	0	0.95	0.00	0.78				
	1	0.41	0.80	0.48				
	2	0.18	1.00	0.33				
Q8 – Difficulty paying attention	-1	1.00	0.00	0.81	0.66	0.16	0.36-0.97	NO
	0	0.95	0.40	0.85				
	1	0.50	0.60	0.52				
	2	0.14	1.00	0.30				
Q9 – Difficulty standing	-1	1.00	0.00	0.81	0.59	0.19	0.22-0.96	NO
	0	0.95	0.40	0.85				
	1	0.41	0.60	0.44				
	2	0.27	0.60	0.33				
Q10 – Difficulty having fun	-1	1.00	0.00	0.81	0.74	0.11	0.53-0.95	YES
	0	0.95	0.20	0.81				
	1	0.59	0.80	0.63				
	2	0.23	1.00	0.37				

Item	Change score (x)	Sensitivity	Specificity	Accuracy	AUC	SE	95% CI	Lower 95% CI ≥0.50
Q11 – Difficulty doing schoolwork	-1	1.00	0.00	0.81	0.59	0.15	0.30-0.88	NO
	0	0.91	0.20	0.78				
	1	0.36	0.80	0.44				
	2	0.18	0.80	0.29				
Q12 – Difficulty walking 1 block	-1	1.00	0.20	0.85	0.91	0.06	0.80-1.00	YES
	0	0.95	0.60	0.89				
	1	0.64	1.00	0.70				
	2	0.36	1.00	0.48				
Q13 – Difficulty running	-1	0.95	0.00	0.78	0.92	0.05	0.82-1.00	YES
	0	0.95	0.60	0.89				
	1	0.77	1.00	0.81				
	2	0.41	1.00	0.52				
Q14 – Keep thinking how much want pain to stop	-1	-	-	-	0.78	0.10	0.59-0.97	YES
	0	1.00	0.00	0.81				
	1	0.73	0.80	0.74				
	2	0.59	0.80	0.63				
Q15 – Afraid pain will get worse	-1	1.00	0.20	0.85	0.74	0.14	0.46-1.00	NO
	0	0.95	0.20	0.81				
	1	0.68	0.80	0.70				
	2	0.41	0.80	0.48				
Q16 – Can't stand it anymore	-1	1.00	0.00	0.81	0.50	0.14	0.24-0.77	NO
	0	0.95	0.00	0.78				
	1	0.59	0.40	0.56				
	2	0.23	0.80	0.33				
Q17 – Feel sad	-1	1.00	0.00	0.81	0.32	0.13	0.06-0.58	NO
	0	0.91	0.00	0.74				
	1	0.32	0.40	0.33				
	2	0.05	0.80	0.19				
Q18 – Feel angry	-1	1.00	0.00	0.81	0.66	0.09	0.49-0.84	NO
	0	0.91	0.20	0.78				
	1	0.27	1.00	0.41				
	2	0.05	1.00	0.22				
Q19 – Feel cheerful	-1	0.91	0.00	0.74	0.35	0.16	0.03-0.67	NO
	0	0.82	0.20	0.70				
	1	0.23	0.40	0.26				
	2	0.14	0.80	0.26				
Q20 – Feel worried	-1	1.00	0.00	0.81	0.55	0.11	0.34-0.77	NO
	0	0.86	0.00	0.70				
	1	0.41	0.80	0.48				
	2	0.05	1.00	0.22				

Sensitivity = probability of having change score of ≥x, given that pain has improved. Specificity = probability of not having change of ≥x, given that pain has not improved. Accuracy = percentage of patients whose change score of ≥x correctly classifies them as better or not better

3.5.8 Summary of responsiveness analyses

Responsiveness at 1-week post-joint injection was demonstrated for 9 of the 20 (45%) SUPER-KIDZ items by at least one statistical method: Q1, Q5, Q8, Q10, Q13-16 (Table 25). As predicted in study hypothesis H2, the SUPER-KIDZ items were more responsive at 2 weeks post-injection with an increase in both the number of responsive questions identified and the magnitude of the results. At 2 weeks post-injection, 16 of 20 (80%) SUPER-KIDZ items (Q1, Q2, Q4, Q5, Q7-17) demonstrated responsiveness based on at least one analysis (Table 26). The most consistently responsive items across all analyses were Q1, Q2, Q4, Q10, and Q12-Q15. Question #5 and #16 were also strongly responsive based on the SRM, Wilcoxon signed rank, and mixed model regression analyses, but not by the ROC curve analysis.

Table 25: Summary of responsiveness results for SUPER-KIDZ items at 1-week post-joint injection (all n=29, improved n=22)

Domain	Item	Question	SRM (all/ improved subjects)	Wilcoxon Signed Rank	Mixed model regression
Pain characteristics	Q1	Current pain	Y/Y	Y	Y
	Q2	Avg pain past week	Y/Y	N	N
	Q3	Pain frequency		N	N
	Q4	Pain duration		N	N
	Q5	# Painful locations	Y/Y	Y	Y
Associated	Q6	Fatigue frequency		N	N
Interference	Q7	Sleeping		N	N
	Q8	Paying attention		Y	Y
	Q9	Standing		N	N
	Q10	Having fun		N	Y
	Q11	Schoolwork		N	N
	Q12	Walk 1 block		N	N
	Q13	Running		Y	Y
	Q14	Want pain to stop		Y	Y
	Q15	Afraid get worse		Y	Y
	Q16	Can't stand it anymore		Y	Y
Emotional	Q17	Feel sad		N	N
	Q18	Feel angry		N	N
	Q19	Feel cheerful		N	N
	Q20	Feel worried		N	N
Totals:			Yes=3	Yes=8	Yes=9

Yes = met pre-specified criteria for responsiveness at 1 week

Table 26: Summary of responsiveness results for SUPER-KIDZ items at 2-weeks post-joint injection (all n=27, improved n=22)

Domain	Item	Question	SRM (all/improved subjects)	Wilcoxon Signed Rank	Mixed model regression	ROC Curve analysis
Pain characteristics	Q1	Current pain	Y/Y	Y	Y	Y
	Q2	Avg pain past week	Y/Y	Y	Y	Y[§]
	Q3	Pain frequency		Y	N	Y
	Q4	Pain duration		Y	Y	Y
	Q5	# Painful locations	Y/Y	Y	Y	N
Associated	Q6	Fatigue frequency		N	N	N
Interference	Q7	Sleeping		Y	Y	N
	Q8	Paying attention		Y	Y	N
	Q9	Standing		Y	Y	N
	Q10	Having fun		Y	Y	Y
	Q11	Schoolwork		Y	Y	N
	Q12	Walk 1 block		Y	Y	Y
	Q13	Running		Y	Y	Y
	Q14	Want pain to stop		Y	Y	Y
	Q15	Afraid get worse		Y	Y	Y[§]
	Q16	Can't stand it anymore		Y	Y	N
Emotional functioning	Q17	Feel sad		Y	Y	N
	Q18	Feel angry		N	N	N
	Q19	Feel cheerful		N	N	N
	Q20	Feel worried		N	N	N
Totals:			Yes=3	Yes=16	Yes=15	Yes=9

Yes = met pre-specified criteria for responsiveness at 2 weeks; [§]AUC≥0.70 but 95% CI crosses 0.50

4 DISCUSSION

4.1 Key findings of the study

This prospective study with repeated measures assesses the test-retest reliability and responsiveness of a new multi-dimensional pain measure, the SUPER-KIDZ tool, in a population of 51 JIA patients aged 8-18 years. This is the first study to examine the measurement properties of a web-based pain measure in children and youth with JIA, and begins to fill a long-standing gap of a validated comprehensive pain assessment measure in paediatric rheumatology. We found that the 3 SUPER-KIDZ subscales have good internal consistency (ordinal α=0.73-0.92). Fifteen of the 20 SUPER-KIDZ items appear to have acceptable test-retest reliability (ICC [2,1] or weighted $\kappa \geq 0.80$) (Table 16), and the sensory questions (Q1-Q5) were the most strongly reliable (reliability coefficients 0.76-0.97). Sixteen of the SUPER-KIDZ items achieved the criteria for responsiveness at 2 weeks by several analytic methods (Table 26). The most strongly responsive questions for the measurement of pain in patients with JIA included most of the sensory items (Q1, Q2, Q4, Q5) as well as many of the interference items (Q10, Q12-Q16). If the final analysis corroborates these interim analyses, it is worthwhile to move forward in investigating construct validity of the tool and the measurement properties of the other versions of the measure (young child and parent-proxy).

4.2 Discussion of study results

4.2.1 Internal consistency of the SUPER-KIDZ tool

The 3 SUPER-KIDZ subscales had acceptable ordinal reliability α parameters. The results are similar to Cronbach α values for other comparable measures, including the PedsQL 'Pain and hurt' subscale (α=0.86) and 'Emotional functioning' subscale (α=0.79) [67], and the Pain Catastrophizing Scale (α=0.68-0.79) [118]. The interference (cognitive-evaluative) subscale had the highest ordinal reliability α (0.92), likely because it has the most items. It is somewhat lower than for the original PROMIS pain interference scale (α=0.96-0.99) [137], which contained 41 items, and did not include cognitive questions.

Question #19 (feeling cheerful) had an ITC of <0.30, suggesting that it is not helpful in discriminating between patients with higher or lower mood from pain. This may be because it

was the only question written in a reverse scoring order (higher score meant more cheerful versus lower score meant less sad/angry/worried for the other affective items), which could be confusing for the children completing the questionnaire. Likewise, it may not be a useful item for assessment of emotional function. Also, the ordinal reliability α parameter increased upon removal of Q19, suggesting that it may not measure the same construct as the other items in the emotional domain. Thus, as written, Q19 may not be suitable for inclusion in the emotional functioning domain of the SUPER-KIDZ tool. The wording and response options may require revision to make it more similar to the other items in the scale (see section 4.3).

4.2.2 Test-retest reliability

It is likely that the majority of the SUPER-KIDZ items are reliable, given that the reliability coefficients for 15 of the 20 items met the hypothesized criteria of ≥0.80 in at least one analysis. Our results provide reasonable evidence for reliability when compared to the test-retest reliability of other pediatric and adult pain measures. The Bath Adolescent Pain Questionnaire (BAPQ) assesses the impact of chronic pain on adolescents, and includes social and physical functioning as well as depression and anxiety subscales. The test-retest reliability correlation coefficients of the BAPQ subscales over a period of 17 days in adolescent rheumatology patients (some with JIA) were 0.79-0.94 [138]. The Pediatric Pain Questionnaire has also been studied in children aged 8-16 years with rheumatic conditions. The pain intensity items (10 cm-VAS) were found to have "moderate stability" (correlations of 0.41 and 0.33 for current pain and worst pain over past week respectively) over a 6-month time frame [63]. Reviews on the psychometric properties of the 10 cm-pain VAS in other pediatric populations report reliability coefficients of 0.58-0.70 [61, 139]. Note that none of the pediatric studies used ICC or kappa parameters to evaluate test-retest reliability thus cannot be directly compared with our results. Childs et al measured the ICC (2,1), SEM and MDC_{95} of the11-point NRS in adults with chronic low back pain (LBP) [140]. In subjects with self-reported 'stable' pain over 1 week, the ICC was 0.61 (0.30-0.77), which is slightly lower than our results of 0.76-0.87 for the 11-point SUPER-KIDZ items Q1 (current pain intensity) and Q2 (average pain intensity over past week) in stable JIA patients. The SEM and MDC_{95} for the 11-point NRS were 1.02 and 1.99 units respectively, which are somewhat smaller than our results for Q1 and Q2. In addition, while the pediatric body location pain map has been tested in children, reliability has not been assessed beyond percent agreement of number of painful sites

(91%) [119, 141]. The ICC of the McGill pain map in adult rheumatology patients was found to be between 0.71-0.84 for number of painful areas [142], which is similar to our findings of ICC between 0.70-0.91 for Q5 (number of painful body locations). Although the above results are difficult to compare with those of the current study given the differing time frames, populations, and reliability statistics used, it appears that our results for test-retest reliability and measurement error are similar or better than those reported in the existing literature.

Within both reliability subgroups, less than half of the subjects reported stability in pain over the corresponding 1-week time period, and a proportion of subjects reported worsening (15-30%) or improvement (24-38%) in pain. It could be that these patients had a change in disease activity over the 1-week study period, although we do not have repeated disease measures to verify this. This finding was not too unexpected, as the fluctuating nature of pain in JIA has been clearly documented in previous studies [28, 63, 65, 142]. For example, the eOuch diary study[65] showed considerable variability in pain intensity, unpleasantness, and interference both within and between days, and that adolescents tended to report higher scores in the mornings. Thus, the time of day that subjects completed the questionnaire may have impacted their responses, although the entries were not time-stamped. The variability in pain may also be due to the previously described findings that the pain experience in JIA depends only partially on disease activity, and is also related to other factors such as cognitive-behavioural, emotional and environmental influences. These predictors were not specifically measured in the current study, but our data supports the importance of assessing these influences as part of a comprehensive pain evaluation. Knowledge of these issues led to a priori selection of the lower criterion of 0.80 for the reliability coefficients of the SUPER-KIDZ pain measure.

Several interesting trends emerged in the test-retest reliability analyses. The questions with the strongest evidence of test-retest reliability were the pain characteristic items (Q1-Q5). In contrast, Q7 (difficulty sleeping), Q10 (difficulty having fun), Q14 (thinking about how much want pain to stop), Q15 (afraid pain would get worse) and Q18 (feeling angry) did not achieve the hypothesized value in any of the analyses. Aside from Q7, these questions have a cognitive or emotional component, and thus may be more prone to fluctuation from moment to moment even with stable pain intensity. In addition, lower kappas were obtained when the analysis was limited to the stable subjects (secondary analysis). These results highlight some of the limitations associated with calculating kappas in small sample sizes. Kappa is an average rate

of agreement and, consequently, a few particularly large agreements or disagreements, especially in a very small sample size, can give misleading results [72]. In addition, a high kappa depends on scores in the cells along the diagonal of a cross-table of the responses, that is, when there is an even distribution of marginals. A homogeneous population, for example arising from clustering at the low or high end of the scale (floor and ceiling effects), gives a skewed marginal distribution. This leads to a higher fraction of chance agreement, and leaves less room for real agreement [74, 143]. Thus, the homogeneity of Q7, Q14 and Q18 may also partially explain the lower reliability values obtained for these items.

The weighted kappa results tended to be higher in the 'injected' group compared with the 'not injected' group. Poor agreement may occur if patients are giving different responses on repeat testing for a systematic reason, for example, due to a regression to the mean phenomenon in the 'not injected' group that was starting with higher pain scores [123]. To support this hypothesis, the test 1 scores tended to be higher in the 'not injected' group compared with the 'injected' group, and the differences between test 2-test 1 scores for the 'not injected' group were consistently negative. Another possibility is that some of the subjects may have had a change in therapy several weeks prior to enrolling in the study, and this was still taking effect, as arthritis medications often take weeks or months to reach maximum effectiveness. Interestingly, Stinson et al [65] also found that levels of pain intensity, interference, unpleasantness and stiffness decreased over the course of one week in adolescents using the eOuch pain diary. Although some of the patients had a change in medication, the difference was not fully explained by this variable. Thus, a combination of factors may be contributing to the lower reliability parameters observed in the 'not injected' group.

4.2.3 Responsiveness

The majority of study subjects reported improvement in pain after intra-articular steroid injection, as expected based on existing evidence [8, 132, 144]. Many of the SUPER-KIDZ items appear to be responsive to this change in pain, given that 16 of the 20 (80%) items met the hypothesized criteria at 2-weeks post-injection in at least one analysis. The consistently least responsive items by all methods included fatigue frequency (Q6), and emotional items Q18 (feeling angry), Q19 (feeling cheerful) and Q20 (feeling worried). Of these, Q18 and Q20 appeared to have floor effects, which may contribute to the poor responsiveness. This

observation may also reflect the fact that individuals are unlikely to experience a profound change in fatigue frequency and mood symptoms in response to a purely medical intervention, and over a short time frame. It might be more relevant to assess responsiveness of these items after a psychological intervention such as CBT. These findings are similar to those reported in a previous study where Stinson et al assessed the responsiveness of the e-Ouch© pain diary in adolescents undergoing joint injection, and found that patients' fatigue score did not significantly change after injection based on insignificant repeated-measures ANOVA and ES of 0.26 at 2 weeks post-injection [65]. Also, the 'psychosocial health' scale of the PedsQL was found to have a lower ES (0.72) than the pain intensity (1.19) and physical function (1.00) scales when measured after medical treatment of the child's rheumatic condition [67]. Thus, it appears that measures of fatigue and emotional function are less reflective of change in pain in patients with JIA undergoing medical treatment compared with sensory and interference items.

As predicted in study hypothesis H2, the SUPER-KIDZ items were more responsive at 2 weeks post-injection compared to 1 week post-injection, with an increase in both the number of responsive questions identified and the magnitude of the results. These results are also consistent with Stinson et al [65] who reported a significant effect of time at 2 weeks post-injection but not at 1-week post-injection for the pain intensity, unpleasantness, and interference scales of the e-Ouch© pain diary. The ES were low-moderate (0.22-0.33) at 1 week post-injection, and moderate-high (0.52-0.71) at 2 weeks post-injection. These findings suggest that an assessment of pain 2 weeks after joint injection is likely to yield a significant change in response if there has been improvement in pain.

At the 2-week time point, the magnitude of change in the pain intensity scores (Q1, Q2) in the primary analysis was approximately 1.7 units on the 11-point scale, which is similar to a study by Brostrom et al who reported a median change of 1.5 units in pain intensity in JIA patients 2 weeks after joint injection [144]. Stinson et al reported a smaller change of 0.7 units in pain intensity 2 weeks post-joint injection [65], which may be due to the lower baseline pain scores in their sample. The changes in score for Q1 and Q2 are smaller than the MDC_{95} of 3-4 units determined through reliability testing. Similarly, the magnitude of change for Q5 (2.3) is considerably smaller than the corresponding MDC_{95} of 6-8 units. Thus, these change scores are within measurement error for this sample. The large measurement error in our sample may be due to several factors including small sample size, instability of the pain construct, and

regression to the mean phenonomenon. Analysis of the full target sample will hopefully reduce the SEM and MDC_{95}. In addition, a formal evaluation of the MCID for the SUPER-KIDZ items would be important to determine how this value compares to the MDC_{95}. This could be done using an anchor-based approach to identify "minimally changed" patients (e.g. patients reporting they are "a little better" on GRCP), as was done by Filocamo et al in their evaluation of the 21-numbered circle VAS in JIA patients [145]. Using this method they found an MCID of -1.1 to -2.2 in slightly improved patients. Revicki et al [146] emphasize the importance of ensuring that the external anchor correlates with the measure under study, and being confident about what constitutes a minimal change in the anchor itself. They note that retrospective self-report may be subject to recall bias, and encourage the use of several independent anchors in several different samples to get a range of MCID values.

As expected, the magnitude of the responsiveness results for the SUPER-KIDZ items differed according to the type of change being examined. For example, larger SRMs were obtained when measuring change in the patients deemed to have improved by the GRCP as compared to the whole joint injection cohort. Similarly, in studies of adults with LBP, the ES of the 11-point pain intensity NRS increased from 1.1 in the whole group to 1.5 when the analysis was limited to patients who self-reported improvement [140, 147]. The ROC curve analysis results also highlight the point that responsiveness depends on the construct of change. Although Q3 (pain frequency) was not found to be responsive by the regression method, it was among the more responsive items identified in the ROC curve analysis. Thus Q3 may be a good item for distinguishing improved from unimproved individuals, despite having borderline internal responsiveness. Conversely, Q5 (number of painful body locations) was found to be responsive using the SRM, Wilcoxon Rank Sum and regression but not by the ROC curve analysis. Husted et al also found differing results for internal and external responsiveness of physical function measures in psoriatic arthritis patients [87], underscoring the importance of specifying the construct of change being quantified when designing a responsiveness study. Thus, our results are consistent with the concept that the magnitude of the responsiveness parameter depends on the construct of change under study [93].

4.3 Suggestions for modification of the SUPER-KIDZ tool

Based on the results of this study, it would be reasonable to consider modifying certain of the SUPER-KIDZ items found to have less acceptable measurement properties. Question #19 (feeling cheerful) correlated poorly with the other items in the emotional subscale, thus this item should likely be modified. Firstly, the scale could be changed so that it increases in the same direction as the other items or it should be separated from the other items and given a distinct scale. Secondly, we could consider using different wording to describe the underlying construct (e.g. "happy", "glad"). As noted above, Q19 along with Q6 (fatigue frequency), Q18 (feeling angry), and Q20 (feeling worried) do not appear responsive to change in pain after a joint injection. In order to limit the impact of factors other than pain on mood score, it may be worthwhile to modify the wording of the instructions to specify: "how often you have felt this way *because of pain*?" These items should also be re-evaluated after a cognitive intervention for pain to determine responsiveness in this scenario.

Items #7 (difficulty sleeping), #10 (difficulty having fun), #14 (thinking about how much want pain to stop), #15 (afraid pain would get worse) and #18 (feeling angry), did not meet the criteria for test-retest reliability. However, all of the weighted kappas were between 0.60-0.77, thus they are reasonably reliable given the instability of the pain construct. Floor and ceiling effects may contribute in part to the low reliability coefficients for Q7, Q14 and Q18. For these, and the other items with floor (Q8, Q11, Q17) and ceiling (Q3, Q4, Q13) effects, it may be worthwhile to consider changing the corresponding scales to broaden the response options. For example, instead of "never, almost never, sometimes, almost always, always", we could use " never, rarely (<10% of the time), occasionally (30% of the time), sometimes (50% of the time), frequently (70% of the time), usually (90% of the time) and always" [148] to increase the heterogeneity of responses.

Another point of discussion is to determine the appropriate and most meaningful scoring algorithm for the SUPER-KIDZ tool. It is likely that each domain should be scored separately since the various aspects of the pain construct may change differently over time, and may require specifically targeted interventions (e.g. physical therapy versus CBT). However, it would be informative to determine the inter-correlations between all of the SUPER-KIDZ items in order to identify which items are highly correlated (interchangeable), and therefore

measuring related aspects of the pain construct [149]. In contrast, identifying items that are less highly correlated would argue that they are measuring unique aspects of the construct and should be assessed separately.

4.4 Limitations

As with all studies, the current study has limitations that are important to address. The main issues are of sample size, combining groups for the reliability analysis, the use of the GRCP as an external criterion, and generalizability of the results.

Given the interim nature of the analysis, one of the main limitations of the study was the relatively small sample size for each analysis group (n=26-33), especially the secondary analyses (n=11-22), which may limit the power to detect test-retest reliability and responsiveness. In particular, the ROC curve analysis was the most underpowered because only 5 of 27 patients had not improved at 2 weeks post-injection, which is a small comparison group. In order to identify an AUC of 0.70 or larger, even with a large SE of 0.10, 15 subjects per group are required (and n=55 per group is required for an SE of 0.05) [150]. As such, the results obtained in this analysis are imprecise estimates for responsiveness. In addition, the SRM for Q5 (number of painful locations) did not increase in the "improved" group at either the 1- or 2- week time points. This was due to an increase in the SD of the difference scores relative to the incremental increase in the mean difference scores, likely a result of the large range of scores for Q5 and the reduced sample size. Nonetheless, despite the small sample size, the majority of the SUPER-KIDZ items achieved the criteria for test-retest reliability and responsiveness by at least one analysis method. Fortunately, recruitment is ongoing and the larger samples will enable further refinement and strengthening of these results.

Another limitation was that subjects from study groups A and B were combined for the reliability analysis. It was felt that patients who had not undergone joint injection in the recent past could be combined together (i.e. 'not injected' group), although it is also possible that they do not represent the same type of patient and perhaps would behave differently in the reliability analysis. However, the subject characteristics and baseline scores were similar between the groups, and the proportion of patients reporting stable pain was also the same as the other reliability group. It is interesting that the reliability coefficients tended to be lower from this subgroup, indicating that they were somehow less stable or perhaps more homogeneous than

the ‘injected’ subgroup. If anything, this trend biased our results conservatively, avoiding a type I error.

An additional issue to discuss is the reliance on the 5-point GRCP as an external criterion to identify stable and improved patients for various parts of the analysis, as there are several weaknesses and limitations associated with GRCS measures [103, 107-109]. Firstly, specific to the current study, the wording of the GRCP made reference to the past 1-week, and therefore required some extrapolation of the results for the change between 2 weeks post-injection and baseline status. It would have also been better to specify an anchoring time point (i.e. ‘compared to how you felt before the joint injection…’) to help orient the patients. A pain GRCS has not been formally validated in the paediatric age group [21]. Juniper et al used a 15-point GRCS in paediatric patients with asthma and found that a notable problem was that younger children had difficulty with the concept of time specification of “the last week” [105]. Even in adult studies, patients tend to have difficulty judging whether they have changed compared to a past time point, and GRCS ratings tend to correlate much more with *current* health state rather than the *change* in health state [107-109]. Also, like any retrospective measure, GRCS is subject to recall bias. Miron-Shatz et al have shown that individuals tend to report unpleasant feelings more strongly than they were experienced (‘Memory-Experience Gap’) [151]. As a way of accounting for the possible inaccuracy of the GRCP, the analysis was performed both in the whole group and then also in the subjects who reported either stability (reliability) or improvement (responsiveness).

Our study cohort included a relatively large proportion of RF positive polyarticular patients, possibly due to the focus on older patients. RF positive patients tend to be more severely affected than the other subtypes [14, 152], which could potentially lead to less stability in pain and lower reliability results. There is no evidence that RF positive patients respond differently to steroid injections compared with the other subtypes [153], so responsiveness is unlikely to be affected. Overall, the disease activity level of the cohort based on number of joints and PGA was in the mild-to-moderate range and the study cohort appeared reflective of the disease characteristics of JIA reported in other Canadian studies [4, 65].

4.5 Conclusions and future directions

The results of this study have met the hypothesized criteria for acceptable test-retest reliability and responsiveness for the majority of the SUPER-KIDZ pain measure items in a cohort of children and youth with JIA aged 8-18 years. In addition, the subscales of the SUPER-KIDZ tool have good internal consistency. The results found in the current study were consistent with or better than those reported for other pain measures in the pediatric and adult populations. This study represents another step in the fulfillment of a long-standing gap in the comprehensive assessment of pain in patients with JIA.

The next steps in the validation of the SUPER-KIDZ tool include establishing the construct validity of the SUPER-KIDZ pain measure, as well as validating the younger patient and parent-proxy versions, and French-language versions; studies which are currently underway. In addition, it will be important to potentially modify some of the items to increase the accuracy of responses and to determine which items are inter-correlated to guide the scoring the SUPER-KIDZ pain measure. It would also be helpful to determine the MCID of the SUPER-KIDZ items using a valid external anchor(s) in order to know if they are greater than the corresponding MDC_{95}, which would be important for the interpretation of change scores.

The SUPER-KIDZ tool is intended for use in the everyday clinic setting. It is likely to be a clinically useful measure as it is quick and easy to use, and its measurement properties are being evaluated in a real world observational setting. If the SUPER-KIDZ pain measure is validated, it can be implemented as a standard measure administered at every clinic visit. This consistent and comprehensive approach to pain assessment should help clinicians to more effectively assess, treat, and monitor pain outcomes in the JIA population. The computerized measure allows for real-time result read-outs for the clinicians and patients to use immediately during the clinical visit. It is likely that use of the SUPER-KIDZ pain measure will increase patient and physician communication about pain at the clinical encounter, improve patient engagement in decision-making about their pain treatment plan and ultimately lead to improved child health outcomes (e.g. HRQL).

Following the validation of the SUPER-KIDZ pain measure, future directions for study include: (a) further studies in pain assessment and treatment in patients with JIA and its impact on HRQL; (b) a better understanding of the interrelationship of disease and pain trajectories

over time; (c) the assessment of predictors of pain in children with rheumatic conditions; and (d) the inclusion of a convenient, standardized pain assessment tool in treatment protocols and as a discriminating factor in differential diagnosis.

References

1. Tucker LB, Dancey, P., Oen, K., Huber, A.M., Lagace, C., Espinosa, V.: **Canadian National Surveillance for Juvenile Idiopathic Arthritis**. In: *American College of Rheumatology.* San Francisco, CA: Arthritis and Rheumatism; 2008.

2. Petty RE, Southwood TR, Manners P, Baum J, Glass DN, Goldenberg J, He X, Maldonado-Cocco J, Orozco-Alcala J, Prieur AM *et al*: **International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001**. *J Rheumatol* 2004, **31**(2):390-392.

3. Packham JC, Hall MA: **Long-term follow-up of 246 adults with juvenile idiopathic arthritis: functional outcome**. *Rheumatology (Oxford)* 2002, **41**(12):1428-1435.

4. Oen K, Duffy CM, Tse SM, Ramsey S, Ellsworth J, Chedeville G, Chetaille AL, Saint-Cyr C, Cabral DA, Spiegel LR *et al*: **Early outcomes and improvement of patients with juvenile idiopathic arthritis enrolled in a Canadian multicenter inception cohort**. *Arthritis Care Res (Hoboken)* 2010, **62**(4):527-536.

5. Solari N, Viola S, Pistorio A, Magni-Manzoni S, Vitale R, Ruperto N, Ullmann N, Filocamo G, Martini A, Ravelli A: **Assessing current outcomes of juvenile idiopathic arthritis: a cross-sectional study in a tertiary center sample**. *Arthritis Rheum* 2008, **59**(11):1571-1579.

6. Brabnikova Maresova K: **Secondary osteoporosis in patients with juvenile idiopathic arthritis**. *J Osteoporos* 2011, **2011**:569417.

7. Huber AM, Gaboury I, Cabral DA, Lang B, Ni A, Stephure D, Taback S, Dent P, Ellsworth J, LeBlanc C *et al*: **Prevalent vertebral fractures among children initiating glucocorticoid therapy for the treatment of rheumatic disorders**. *Arthritis Care Res (Hoboken)* 2010, **62**(4):516-526.

8. Bloom BJ, Alario AJ, Miller LC: **Intra-articular corticosteroid therapy for juvenile idiopathic arthritis: report of an experiential cohort and literature review**. *Rheumatol Int* 2010, **31**(6):749-756.

9. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A: **Preliminary definition of improvement in juvenile arthritis**. *Arthritis Rheum* 1997, **40**(7):1202-1209.

10. Takken T, Van Brussel M, Engelbert RH, Van Der Net J, Kuis W, Helders PJ: **Exercise therapy in juvenile idiopathic arthritis: a Cochrane Review**. *Eur J Phys Rehabil Med* 2008, **44**(3):287-297.

11. Philpott J, Houghton K, Luke A: **Physical activity recommendations for children with specific chronic health conditions: Juvenile idiopathic arthritis, hemophilia, asthma and cystic fibrosis**. *Paediatr Child Health*, **15**(4):213-225.

12. Hackett J, Johnson B, Parkin A, Southwood T: **Paediatric rheumatology: clinical practice review. Physiotherapy and occupational therapy for juvenile chronic arthritis: custom and practice in five centres in the UK, USA and Canada**. *Br J Rheumatol* 1996, **35**(7):695-699.

13. Stagi S, Masi L, Capannini S, Cimaz R, Tonini G, Matucci-Cerinic M, de Martino M, Falcini F: **Cross-sectional and longitudinal evaluation of bone mass in children and young adults with juvenile idiopathic arthritis: the role of bone mass determinants in a large cohort of patients**. *J Rheumatol*, **37**(9):1935-1943.

14. Ravelli A, Martini A: **Juvenile idiopathic arthritis**. *Lancet* 2007, **369**(9563):767-778.

15. Taxonomy ITFo. In: *Classification of Chronic Pain, Second Edition.* Edited by Bogduk HMaN. Seattle: IASP Press; 1994: pp 209-214.

16. Palermo T EC, Goldschneider K, McGinn L, Sethna N, Schechter N, Turner H: **Assessment and Management of Children with Chronic Pain: American Pain Society Position Statement**. *American Pain Society* 2012.

17. Anthony KK, Schanberg LE: **Pediatric pain syndromes and management of pain in children and adolescents with rheumatic disease**. *Pediatr Clin North Am* 2005, **52**(2):611-639, vii.

18. Melzack R, Torgerson WS: **On the language of pain**. *Anesthesiology* 1971, **34**(1):50-59.

19. Melzack RC, K.L.: **Sensory motivational and central control determinants of pain: A new conceptual model**: Springfield: Thomas.; 1968.

20. von Baeyer CL: **Children's self-reports of pain intensity: scale selection, limitations and interpretation**. *Pain Res Manag* 2006, **11**(3):157-162.

21. McGrath PJ, Walco GA, Turk DC, Dworkin RH, Brown MT, Davidson K, Eccleston C, Finley GA, Goldschneider K, Haverkos L *et al*: **Core outcome domains and measures for pediatric acute and chronic/recurrent pain clinical trials: PedIMMPACT recommendations**. *J Pain* 2008, **9**(9):771-783.

22. Lovell DJ, Walco GA: **Pain associated with juvenile rheumatoid arthritis**. *Pediatr Clin North Am* 1989, **36**(4):1015-1027.

23. Schanberg LE, Lefebvre JC, Keefe FJ, Kredich DW, Gil KM: **Pain coping and the pain experience in children with juvenile chronic arthritis**. *Pain* 1997, **73**(2):181-189.

24. Scott PJ, Ansell BM, Huskisson EC: **Measurement of pain in juvenile chronic polyarthritis**. *Ann Rheum Dis* 1977, **36**(2):186-187.

25. Sherry DD, Bohnsack J, Salmonson K, Wallace CA, Mellins E: **Painless juvenile rheumatoid arthritis**. *J Pediatr* 1990, **116**(6):921-923.

26. Thompson KL, Varni JW, Hanson V: **Comprehensive assessment of pain in juvenile rheumatoid arthritis: an empirical model**. *J Pediatr Psychol* 1987, **12**(2):241-255.

27. Benestad B, Vinje O, Veierod MB, Vandvik IH: **Quantitative and qualitative assessments of pain in children with juvenile chronic arthritis based on the Norwegian version of the Pediatric Pain Questionnaire**. *Scand J Rheumatol* 1996, **25**(5):293-299.

28. Schanberg LE, Anthony KK, Gil KM, Maurin EC: **Daily pain and symptoms in children with polyarticular arthritis**. *Arthritis Rheum* 2003, **48**(5):1390-1397.

29. Kimura Y, Walco GA, Sugarman E, Conte PM, Schanberg LE: **Treatment of pain in juvenile idiopathic arthritis: a survey of pediatric rheumatologists**. *Arthritis Rheum* 2006, **55**(1):81-85.

30. Kantor TG: **Concepts in pain control**. *Semin Arthritis Rheum* 1989, **18**(4 Suppl 2):94-99.

31. Kuis W, Heijnen CJ, Hogeweg JA, Sinnema G, Helders PJ: **How painful is juvenile chronic arthritis?** *Arch Dis Child* 1997, **77**(5):451-453.

32. Ilowite NT, Walco GA, Pochaczevsky R: **Assessment of pain in patients with juvenile rheumatoid arthritis: relation between pain intensity and degree of joint inflammation**. *Ann Rheum Dis* 1992, **51**(3):343-346.

33. Malleson PN, Oen K, Cabral DA, Petty RE, Rosenberg AM, Cheang M: **Predictors of pain in children with established juvenile rheumatoid arthritis**. *Arthritis Rheum* 2004, **51**(2):222-227.

34. Vandvik IH, Eckblad G: **Relationship between pain, disease severity and psychosocial function in patients with juvenile chronic arthritis (JCA)**. *Scand J Rheumatol* 1990, **19**(4):295-302.

35. Hogeweg JA, Kuis W, Huygen AC, de Jong-de vos van Steenwijk C, Bernards AT, Oostendorp RA, Helders PJ: **The pain threshold in juvenile chronic arthritis**. *Br J Rheumatol* 1995, **34**(1):61-67.

36. Woolf CJ: **Generation of acute pain: central mechanisms**. *Br Med Bull* 1991, **47**(3):523-533.

37. Thastum M, Herlin T, Zachariae R: **Relationship of pain-coping strategies and pain-specific beliefs to pain experience in children with juvenile idiopathic arthritis**. *Arthritis Rheum* 2005, **53**(2):178-184.

38. Thastum M, Herlin T: **Pain-specific beliefs and pain experience in children with juvenile idiopathic arthritis: a longitudinal study**. *J Rheumatol* 2010, **38**(1):155-160.

39. Schanberg LE, Gil KM, Anthony KK, Yow E, Rochon J: **Pain, stiffness, and fatigue in juvenile polyarticular arthritis: contemporaneous stressful events and mood as predictors**. *Arthritis Rheum* 2005, **52**(4):1196-1204.

40. Connelly M, Bromberg MH, Anthony KK, Gil KM, Franks L, Schanberg LE: **Emotion regulation predicts pain and functioning in children with juvenile idiopathic arthritis: An electronic diary study**. *Journal of Pediatric Psychology* 2012, **Vol.37**(1):pp.

41. Schanberg LE, Anthony KK, Gil KM, Lefebvre JC, Kredich DW, Macharoni LM: **Family pain history predicts child health status in children with chronic rheumatic disease**. *Pediatrics* 2001, **108**(3):E47.

42. Connelly M, Anthony KK, Sarniak R, Bromberg MH, Gil KM, Schanberg LE: **Parent pain responses as predictors of daily activities and mood in children with juvenile idiopathic arthritis: the utility of electronic diaries**. *J Pain Symptom Manage*, **39**(3):579-590.

43. Sawyer MG, Whitham JN, Roberton DM, Taplin JE, Varni JW, Baghurst PA: **The relationship between health-related quality of life, pain and coping strategies in juvenile idiopathic arthritis**. *Rheumatology (Oxford)* 2004, **43**(3):325-330.

44. Oliveira S, Ravelli A, Pistorio A, Castell E, Malattia C, Prieur AM, Saad-Magalhaes C, Murray KJ, Bae SC, Joos R *et al*: **Proxy-reported health-related quality of life of patients with juvenile idiopathic arthritis: the Pediatric Rheumatology International Trials Organization multinational quality of life cohort study**. *Arthritis Rheum* 2007, **57**(1):35-43.

45. Gutierrez-Suarez R, Pistorio A, Cespedes Cruz A, Norambuena X, Flato B, Rumba I, Harjacek M, Nielsen S, Susic G, Mihaylova D *et al*: **Health-related quality of life of patients with juvenile idiopathic arthritis coming from 3 different geographic areas. The PRINTO multinational quality of life cohort study**. *Rheumatology (Oxford)* 2007, **46**(2):314-320.

46. Seid M, Opipari L, Huang B, Brunner HI, Lovell DJ: **Disease control and health-related quality of life in juvenile idiopathic arthritis**. *Arthritis Rheum* 2009, **61**(3):393-399.

47. Haverman L, Grootenhuis MA, van den Berg JM, van Veenendaal M, Dolman KM, Swart JF, Kuijpers TW, van Rossum MA: **Predictors of health-related quality of life in children and adolescents with juvenile idiopathic arthritis: results from a Web-based survey**. *Arthritis Care Res (Hoboken)* 2012, **64**(5):694-703.

48. Feinstein AB, Forman EM, Masuda A, Cohen LL, Herbert JD, Moorthy LN, Goldsmith DP: **Pain intensity, psychological inflexibility, and acceptance of pain as predictors of functioning in adolescents with juvenile idiopathic arthritis: a preliminary investigation**. *J Clin Psychol Med Settings* 2011, **18**(3):291-298.

49. Margetic B, Aukst-Margetic B, Bilic E, Jelusic M, Tambic Bukovac L: **Depression, anxiety and pain in children with juvenile idiopathic arthritis (JIA)**. *Eur Psychiatry* 2005, **20**(3):274-276.

50. Butbul Aviel Y, Stremler R, Benseler SM, Cameron B, Laxer RM, Ota S, Schneider R, Spiegel L, Stinson JN, Tse SM *et al*: **Sleep and fatigue and the relationship to pain, disease activity and quality of life in juvenile idiopathic arthritis and juvenile dermatomyositis**. *Rheumatology (Oxford)* 2011, **50**(11):2051-2060.

51. Bloom BJ, Owens JA, McGuinn M, Nobile C, Schaeffer L, Alario AJ: **Sleep and its relationship to pain, dysfunction, and disease activity in juvenile rheumatoid arthritis**. *J Rheumatol* 2002, **29**(1):169-173.

52. Dhanani S, Quenneville J, Perron M, Abdolell M, Feldman BM: **Minimal difference in pain associated with change in quality of life in children with rheumatic disease**. *Arthritis Rheum* 2002, **47**(5):501-505.

53. Tong A, Jones J, Craig JC, Singh-Grewal D: **Children's experiences of living with juvenile idiopathic arthritis: a thematic synthesis of qualitative studies**. *Arthritis Care Res (Hoboken)* 2012, **64**(9):1392-1404.

54. Stinson JN, Luca NJ, Jibb LA: **Assessment and management of pain in juvenile idiopathic arthritis**. *Pain Res Manag* 2012, **17**(6):391-396.

55. Connelly M, Schanberg L: **Opioid therapy for the treatment of refractory pain in children with juvenile rheumatoid arthritis**. *Nat Clin Pract Rheumatol* 2006, **2**(12):636-637.

56. Kimura Y, Walco GA: **Treatment of chronic pain in pediatric rheumatic disease**. *Nat Clin Pract Rheumatol* 2007, **3**(4):210-218.

57. Walco GA, Varni JW, Ilowite NT: **Cognitive-behavioral pain management in children with juvenile rheumatoid arthritis**. *Pediatrics* 1992, **89**(6 Pt 1):1075-1079.

58. Kirshner B, Guyatt G: **A methodological framework for assessing health indices**. *J Chronic Dis* 1985, **38**(1):27-36.

59. Beales JG, Keen JH, Holt PJ: **The child's perception of the disease and the experience of pain in juvenile chronic arthritis**. *J Rheumatol* 1983, **10**(1):61-65.

60. Kuis W, Heijnen CJ, Sinnema G, Kavelaars A, van der Net J, Helders PJ: **Pain in childhood rheumatic arthritis**. *Baillieres Clin Rheumatol* 1998, **12**(2):229-244.

61. Stinson JN, Kavanagh T, Yamada J, Gill N, Stevens B: **Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trials in children and adolescents**. *Pain* 2006, **125**(1-2):143-157.

62. Varni JW, Thompson, K.L.: **The Varni/Thompson Pediatric Pain Quesionnaire**. *Unpublished manuscript* 1985.

63. Gragg RA, Rapoff MA, Danovsky MB, Lindsley CB, Varni JW, Waldron SA, Bernstein BH: **Assessing chronic musculoskeletal pain associated with rheumatic disease: further validation of the pediatric pain questionnaire**. *J Pediatr Psychol* 1996, **21**(2):237-250.

64. Stinson JN, Petroz GC, Tait G, Feldman BM, Streiner D, McGrath PJ, Stevens BJ: **e-Ouch: usability testing of an electronic chronic pain diary for adolescents with arthritis**. *Clin J Pain* 2006, **22**(3):295-305.

65. Stinson JN, Stevens BJ, Feldman BM, Streiner D, McGrath PJ, Dupuis A, Gill N, Petroz GC: **Construct validity of a multidimensional electronic pain diary for adolescents with arthritis**. *Pain* 2008, **136**(3):281-292.

66. Lootens CC, Rapoff MA: **Measures of pediatric pain: 21-numbered circle Visual Analog Scale (VAS), E-Ouch Electronic Pain Diary, Oucher, Pain Behavior Observation Method, Pediatric Pain Assessment Tool (PPAT), and Pediatric Pain Questionnaire (PPQ)**. *Arthritis Care Res (Hoboken)* 2011, **63 Suppl 11**:S253-262.

67. Varni JW, Seid M, Smith Knight T, Burwinkle T, Brown J, Szer IS: **The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory Generic Core Scales and Rheumatology Module**. *Arthritis Rheum* 2002, **46**(3):714-725.

68. Singh G, Athreya BH, Fries JF, Goldsmith DP: **Measurement of health status in children with juvenile rheumatoid arthritis**. *Arthritis Rheum* 1994, **37**(12):1761-1769.

69. Thomson W: **Electrical Units of Measurement**. *Popular Lectures* 1883, **1**.

70. Guyatt GH, Bombardier C, Tugwell PX: **Measuring disease-specific quality of life in clinical trials**. *CMAJ* 1986, **134**(8):889-895.

71. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC: **The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes**. *J Clin Epidemiol* 2010, **63**(7):737-745.

72. Portney LF WM: **Foundation of clinical research:applications to practice, 2nd edition. Chapter 26: Statistical Measures of Reliability**. Norwich, Con: Appleton and Lange; 2000.

73. Nunnally J, Bernstein, IH: **Psychometric Theory 3rd Ed.**, 3rd edn. New York: McGraw-Hill; 1994.

74. de Vet H, Terwee, CB, Mokkink, LB, Knol DL: **Measurement in Medicine: A Practical Guide**. Cambridge, UK: Cambridge University Press; 2011.

75. Gadermann A.M. GM, Zumbo B.D.: **Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide**. In: *Practical Assessment, Research & Evaluation.* vol. 17; 2012: 1-13.

76. Weir JP: **Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM**. *J Strength Cond Res* 2005, **19**(1):231-240.

77. Shrout PE, Fleiss JL: **Intraclass correlations: uses in assessing rater reliability**. *Psychol Bull* 1979, **86**(2):420-428.

78. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC: **Quality criteria were proposed for measurement properties of health status questionnaires**. *J Clin Epidemiol* 2007, **60**(1):34-42.

79. Spadoni GF, Stratford PW, Solomon PE, Wishart LR: **The evaluation of change in pain intensity: a comparison of the P4 and single-item numeric pain rating scales**. *J Orthop Sports Phys Ther* 2004, **34**(4):187-193.

80. Landis JR, Koch GG: **The measurement of observer agreement for categorical data**. *Biometrics* 1977, **33**(1):159-174.

81. Christensen L, Mendoza, JL: **A method of assessing change in a single subject: an alteration of the RC index**. *Behavior Therapy* 1986, **17**(3):305-308.

82. Liang MH: **Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments**. *Med Care* 2000, **38**(9 Suppl):II84-90.

83. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM: **On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation**. *Qual Life Res* 2003, **12**(4):349-362.

84. Beaton DE, Bombardier C, Katz JN, Wright JG: **A taxonomy for responsiveness**. *J Clin Epidemiol* 2001, **54**(12):1204-1217.

85. de Bruin AF, Diederiks JP, de Witte LP, Stevens FC, Philipsen H: **Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP68**. *J Clin Epidemiol* 1997, **50**(5):529-540.

86. Beaton DE: **Understanding the relevance of measured change through studies of responsiveness**. *Spine (Phila Pa 1976)* 2000, **25**(24):3192-3199.

87. Husted JA, Cook RJ, Farewell VT, Gladman DD: **Methods for assessing responsiveness: a critical review and recommendations**. *J Clin Epidemiol* 2000, **53**(5):459-468.

88. Wright JG, Young NL: **A comparison of different indices of responsiveness**. *J Clin Epidemiol* 1997, **50**(3):239-246.

89. Streiner DL, Norman, G. R.: **Health Measurement Scales: A Practical Guide to Their Development and Use, 3rd Edition**: Oxford University Press; 2003.

90. Guyatt G, Walter S, Norman G: **Measuring change over time: assessing the usefulness of evaluative instruments**. *J Chronic Dis* 1987, **40**(2):171-178.

91. Kazis LE, Anderson JJ, Meenan RF: **Effect sizes for interpreting changes in health status**. *Med Care* 1989, **27**(3 Suppl):S178-189.

92. Liang MH, Fossel AH, Larson MG: **Comparisons of five health status instruments for orthopedic evaluation**. *Med Care* 1990, **28**(7):632-642.

93. 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**(2):128-146.

94. Stratford PW, Binkley JM, Riddle DL: **Health status measures: strategies and analytic methods for assessing change scores**. *Phys Ther* 1996, **76**(10):1109-1123.

95. Conner-Spady B, Cumming C, Nabholtz JM, Jacobs P, Stewart D: **Responsiveness of the EuroQol in breast cancer patients undergoing high dose chemotherapy**. *Qual Life Res* 2001, **10**(6):479-486.

96. Arnold SH, McEwen IR: **Item test-retest reliability and responsiveness of the School Outcomes Measure (SOM)**. *Phys Occup Ther Pediatr* 2008, **28**(1):59-77.

97. King M, Dobson A: **Estimating the responsiveness of an instrument using more than two repeated measures**. *Biometrics* 2000, **56**(4):1197-1203.

98. Bell ML, Kenward MG, Fairclough DL, Horton NJ: **Differential dropout and bias in randomised controlled trials: when it matters and when it may not**. *BMJ* 2013, **346**:e8668.

99. Deyo RA, Centor RM: **Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance**. *J Chronic Dis* 1986, **39**(11):897-906.

100. Krahn M, Bremner KE, Tomlinson G, Ritvo P, Irvine J, Naglie G: **Responsiveness of disease-specific and generic utility instruments in prostate cancer patients**. *Qual Life Res* 2007, **16**(3):509-522.

101. Horng YS, Lin MC, Feng CT, Huang CH, Wu HC, Wang JD: **Responsiveness of the Michigan Hand Outcomes Questionnaire and the Disabilities of the Arm, Shoulder, and Hand questionnaire in patients with hand injury**. *J Hand Surg Am* 2010, **35**(3):430-436.

102. Jaeschke R, Singer J, Guyatt GH: **Measurement of health status. Ascertaining the minimal clinically important difference**. *Control Clin Trials* 1989, **10**(4):407-415.

103. Kamper SJ, Maher CG, Mackay G: **Global rating of change scales: a review of strengths and weaknesses and considerations for design**. *J Man Manip Ther* 2009, **17**(3):163-170.

104. Preston CC, Colman AM: **Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences**. *Acta Psychol (Amst)* 2000, **104**(1):1-15.

105. Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M: **Measuring quality of life in children with asthma**. *Qual Life Res* 1996, **5**(1):35-46.

106. Costa LO, Maher CG, Latimer J, Ferreira PH, Ferreira ML, Pozzi GC, Freitas LM: **Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil: which one is the best?** *Spine (Phila Pa 1976)* 2008, **33**(22):2459-2463.

107. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting G: **Methods to explain the clinical significance of health status measures**. *Mayo Clin Proc* 2002, **77**(4):371-383.

108. Guyatt GH, Norman GR, Juniper EF, Griffith LE: **A critical look at transition ratings**. *J Clin Epidemiol* 2002, **55**(9):900-908.

109. Norman GR, Stratford P, Regehr G: **Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach**. *J Clin Epidemiol* 1997, **50**(8):869-879.

110. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N *et al*: **Core outcome measures for chronic pain clinical trials: IMMPACT recommendations**. *Pain* 2005, **113**(1-2):9-19.

111. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, Haythornthwaite JA, Jensen MP, Kerns RD, Ader DN *et al*: **Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations**. *J Pain* 2008, **9**(2):105-121.

112. Fritz JM, Piva SR: **Physical impairment index: reliability, validity, and responsiveness in patients with acute low back pain**. *Spine (Phila Pa 1976)* 2003, **28**(11):1189-1194.

113. Farrar JT, Young JP, Jr., LaMoreaux L, Werth JL, Poole RM: **Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale**. *Pain* 2001, **94**(2):149-158.

114. Miller GA: **The magical number seven plus or minus two: some limits on our capacity for processing information**. *Psychol Rev* 1956, **63**(2):81-97.

115. Powell CV, Kelly AM, Williams A: **Determining the minimum clinically significant difference in visual analog pain score for children**. *Ann Emerg Med* 2001, **37**(1):28-31.

116. Stinson JN, Connelly M, Jibb LA, Schanberg LE, Walco G, Spiegel LR, Tse SM, Chalom EC, Chira P, Rapoff M: **Developing a standardized approach to the assessment of pain in children and youth presenting to pediatric rheumatology providers: a Delphi survey and consensus conference process followed by feasibility testing**. *Pediatr Rheumatol Online J* 2012, **10**(1):7.

117. Varni JW, Stucky BD, Thissen D, Dewitt EM, Irwin DE, Lai JS, Yeatts K, Dewalt DA: **PROMIS Pediatric Pain Interference Scale: an item response theory analysis of the pediatric pain item bank**. *J Pain* 2010, **11**(11):1109-1119.

118. Crombez G, Bijttebier P, Eccleston C, Mascagni T, Mertens G, Goubert L, Verstraeten K: **The child version of the pain catastrophizing scale (PCS-C): a preliminary validation**. *Pain* 2003, **104**(3):639-646.

119. von Baeyer CL, Lin V, Seidman LC, Tsao JC, Zeltzer LK: **Pain charts (body maps or manikins) in assessment of the location of pediatric pain**. *Pain Manag* 2011, **1**(1):61-68.

120. Felson DT, Anderson JJ: **Methodological and statistical approaches to criteria development in rheumatic diseases**. *Baillieres Clin Rheumatol* 1995, **9**(2):253-266.

121. Lovell DJ, Passo MH, Beukelman T, Bowyer SL, Gottlieb BS, Henrickson M, Ilowite NT, Kimura Y, DeWitt EM, Segerman J *et al*: **Measuring process of arthritis care: a proposed set of quality measures for the process of care in juvenile idiopathic arthritis**. *Arthritis Care Res (Hoboken)* 2010, **63**(1):10-16.

122. Connelly M, von Baeyer CL, Stinson J, Schanberg LE: **Pain as a quality of care measure in juvenile idiopathic arthritis: one step forward, but is it the best foot? Comment on the article by Lovell et al**. *Arthritis Care Res (Hoboken)* 2011, **63**(9):1352-1353; author reply 1353.

123. Barnett AG, van der Pols JC, Dobson AJ: **Regression to the mean: what it is and how to deal with it**. *Int J Epidemiol* 2005, **34**(1):215-220.

124. Walter SD, Eliasziw M, Donner A: **Sample size and optimal designs for reliability studies**. *Stat Med* 1998, **17**(1):101-110.

125. PASS11: **PASS11 Software**. In.; 2011.

126. **Explanation of Wilcoxon PSRT for non parametric paired differences** [http://www.stattools.net/Wilcoxon_Exp.php]

127. **Tables of sample size for paired difference** [http://www.stattools.net/SSizpairedt_Tab.php]

128. Ruperto N, Pistorio A, Ravelli A, Hasija R, Guseinova D, Filocamo G, Demirkaya E, Malattia C, Martini A: **Criteria to define response to therapy in paediatric rheumatic diseases**. *Eur J Clin Pharmacol* 2010, **67 Suppl 1**:125-131.

129. Perneger TV: **What's wrong with Bonferroni adjustments**. *BMJ* 1998, **316**(7139):1236-1238.

130. McGraw K, Wong SP: **Forming inferences about some intraclass correlation coefficients**. *Psych Methods* 1996, **1**(1):30-46.

131. Inc. SI: **SAS/Stat(R) 9.2 User's Guide**. In., 2nd Edition edn. Cary NC:: SAS Institute Inc.; 2009.

132. Cheng OT, Souzdalnitski D, Vrooman B, Cheng J: **Evidence-based knee injections for the management of arthritis**. *Pain Med* 2012, **13**(6):740-753.

133. Beaton DE, Hogg-Johnson S, Bombardier C: **Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders**. *J Clin Epidemiol* 1997, **50**(1):79-93.

134. Beaton D, Richards RR: **Assessing the reliability and responsiveness of 5 shoulder questionnaires**. *J Shoulder Elbow Surg* 1998, **7**(6):565-572.

135. **SAS Data Analysis Examples. Robust Regression** [http://www.ats.ucla.edu/stat/sas/notes2/]

136. Rosman AS, Korsten MA: **Application of summary receiver operating characteristics (sROC) analysis to diagnostic clinical testing**. *Adv Med Sci* 2007, **52**:76-82.

137. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, Cella D, Rothrock N, Keefe F, Callahan L *et al*: **Development of a PROMIS item bank to measure pain interference**. *Pain* 2010, **150**(1):173-182.

138. Eccleston C, Jordan A, McCracken LM, Sleed M, Connell H, Clinch J: **The Bath Adolescent Pain Questionnaire (BAPQ): development and preliminary psychometric evaluation of an instrument to assess the impact of chronic pain on adolescents**. *Pain* 2005, **118**(1-2):263-270.

139. Cohen LL, Lemanek K, Blount RL, Dahlquist LM, Lim CS, Palermo TM, McKenna KD, Weiss KE: **Evidence-based assessment of pediatric pain**. *J Pediatr Psychol* 2008, **33**(9):939-955; discussion 956-937.

140. Childs JD, Piva SR, Fritz JM: **Responsiveness of the numeric pain rating scale in patients with low back pain**. *Spine (Phila Pa 1976)* 2005, **30**(11):1331-1334.

141. Savedra MC, Tesler MD, Holzemer WL, Wilkie DJ, Ward JA: **Pain location: validity and reliability of body outline markings by hospitalized children and adolescents**. *Res Nurs Health* 1989, **12**(5):307-314.

142. Escalante A, Lichtenstein MJ, Lawrence VA, Roberson M, Hazuda HP: **Where does it hurt? Stability of recordings of pain location using the McGill Pain Map**. *J Rheumatol* 1996, **23**(10):1788-1793.

143. von Eye A, von Eye M: **On the marginal dependency of Cohen's**. *European Psychologist* 2008, **Vol.13**(4):pp.

144. Brostrom E, Hagelberg S, Haglund-Akerlind Y: **Effect of joint injections in children with juvenile idiopathic arthritis: evaluation by 3D-gait analysis**. *Acta Paediatr* 2004, **93**(7):906-910.

145. Filocamo G, Davi S, Pistorio A, Bertamino M, Ruperto N, Lattanzi B, Consolaro A, Magni-Manzoni S, Galasso R, Varnier GC *et al*: **Evaluation of 21-numbered circle and 10-centimeter horizontal line visual analog scales for physician and parent subjective ratings in juvenile idiopathic arthritis**. *J Rheumatol* 2010, **37**(7):1534-1541.

146. Revicki D, Hays RD, Cella D, Sloan J: **Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes**. *J Clin Epidemiol* 2008, **61**(2):102-109.

147. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N: **Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients**. *BMC Musculoskelet Disord* 2006, **7**:82.

148. Vagias WM: **Likert-type scale response anchors**. Clemson University: Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management; 2006.

149. Atkinson MJ, Lennox RD: **Extending basic principles of measurement models to the design and validation of Patient Reported Outcomes**. *Health Qual Life Outcomes* 2006, **4**:65.

150. Hanley J, McNeil BJ: **The meaning and use of the Area Under a Receiver Operating Characteristic (ROC) curve**. *Radiology* 1982, **143**:29-36.

151. Miron-Shatz T, Stone A, Kahneman D: **Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap?** *Emotion* 2009, **Vol.9**(6):pp.

152. Flato B, Lien G, Smerdel A, Vinje O, Dale K, Johnston V, Sorskaar D, Moum T, Ploski R, Forre O: **Prognostic factors in juvenile rheumatoid arthritis: a case-control study revealing early predictors and outcome after 14.9 years**. *J Rheumatol* 2003, **30**(2):386-393.

153. Ravelli A, Manzoni SM, Viola S, Pistorio A, Ruperto N, Martini A: **Factors affecting the efficacy of intraarticular corticosteroid injection of knees in juvenile idiopathic arthritis**. *J Rheumatol* 2001, **28**(9):2100-2102.

Appendices

1. SUPER-KIDZ pain tool

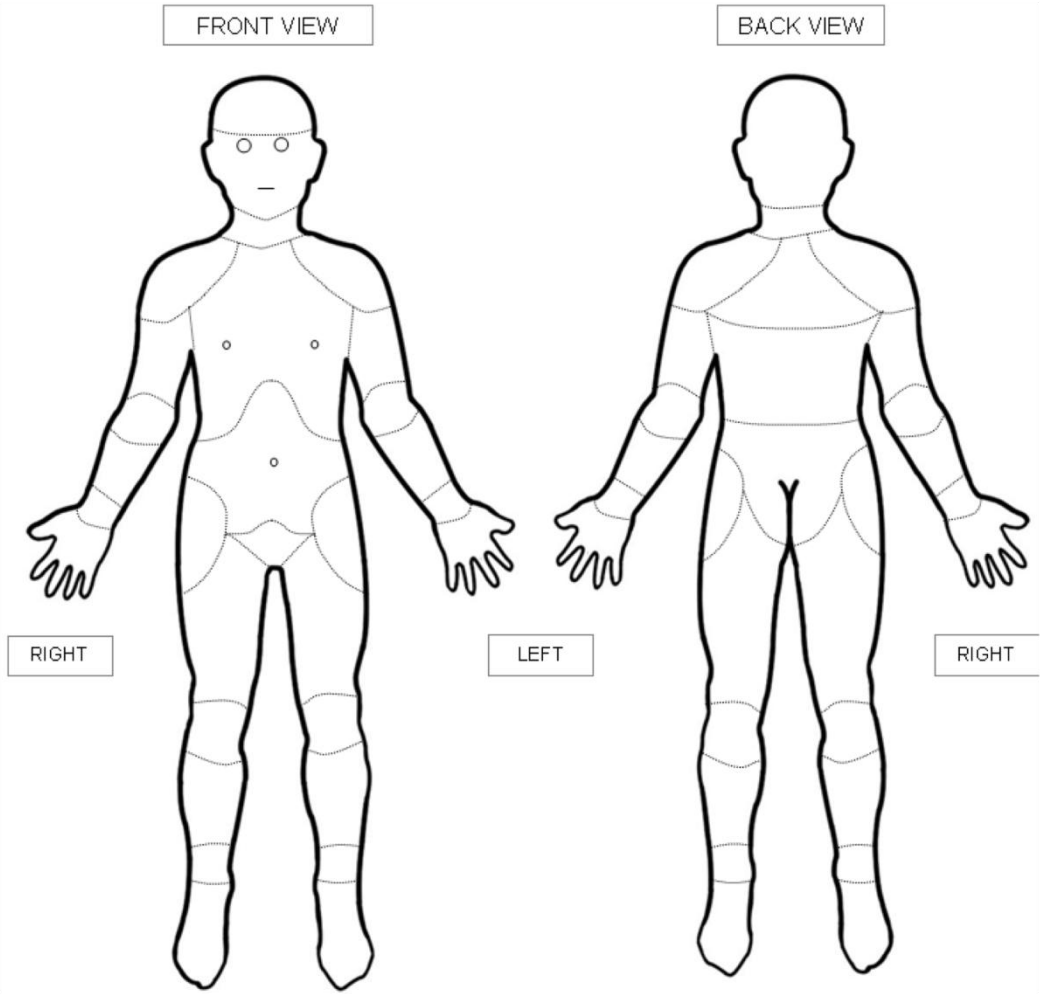
Child 4-8 yrs self-report: Pain intensity and location only

SUPER-KIDZ

Recommended Self-Report Version for Children Aged <8 Years

(1) These faces show how much something can hurt. This first face shows no pain. The faces show more and more pain up to the last face – it shows very much pain. Select the face that shows how much you hurt right now.

(2) Select all the parts of your body where you have had pain in the past 7 days.



Thank you for answering these questions.

Version Date: 18.03.10

Child 4-8 years: Parent proxy-report

SUPER-KIDZ

Recommended Parent-Report Version for Children Aged < 8 Year

Instructions: Please answer the following questions about your child's pain from *your* view.

(1) How much pain do you think your child has *right now*? Check the box below the number that best describes your child's level of pain right now, with 0 being "no pain" and 10 being "most pain possible."

No pain	0	1	2	3	4	5	6	7	8	9	10	Most pain possible
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

(2) If your child had pain *in the past 7 days*, how much did it usually hurt? Check the box below the number that best describes your child's usual level of pain during *the past 7 days*, with 0 being "no pain" and 10 being "most pain possible."

No pain	0	1	2	3	4	5	6	7	8	9	10	Most pain possible
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

(3) On how many days did your child have pain *in the past 7 days?*

- ☐ 1 day
- ☐ 2 days
- ☐3 days
- ☐ 4 days
- ☐ 5 days
- ☐6 days
- ☐ Every day

(4) If your child had pain *in the past 7 days*, how long did the pain usually last?

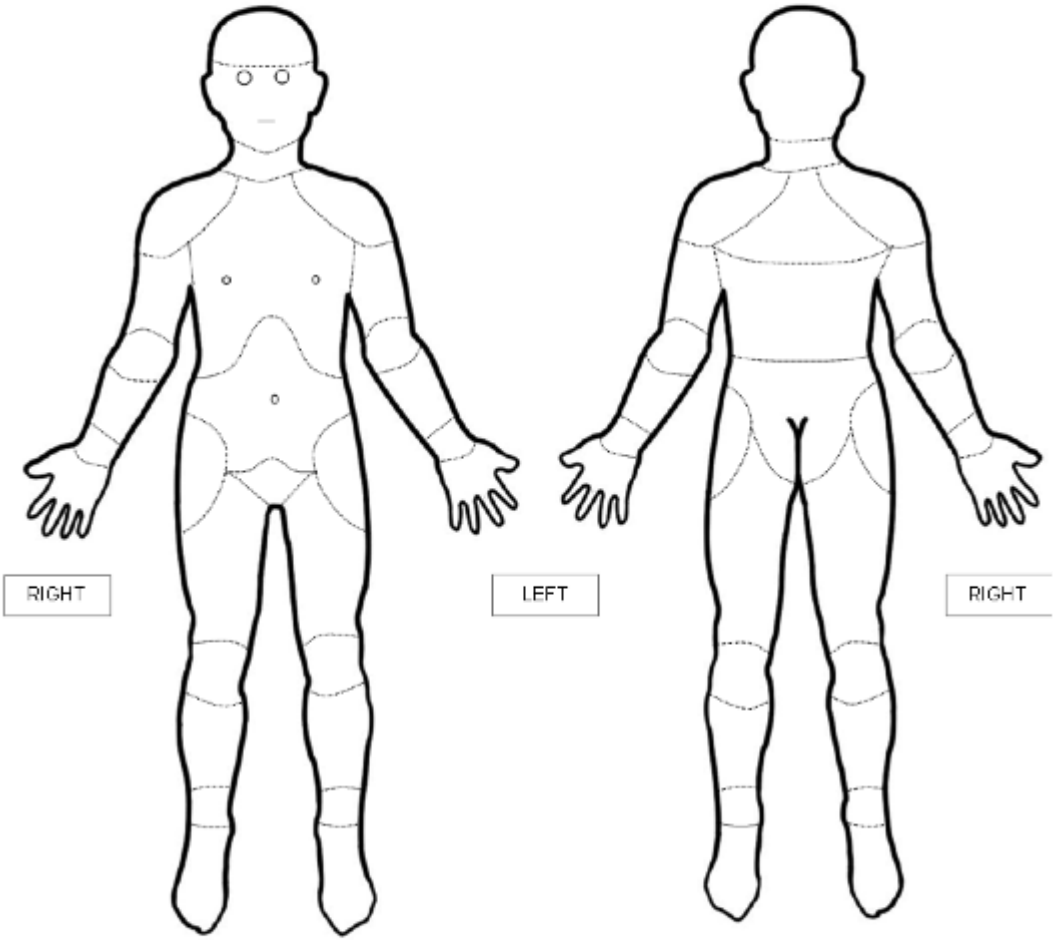
- ☐ Less than 30 minutes
- ☐ About 1 hour
- ☐ Between 1 and 3 hours
- ☐ About half the day
- ☐ All day or longer
- ☐ No pain in the past 7 days

Version Date: 18.03.10

(5) Select all the parts of the body where your child has had pain *in the past 7 days*.

FRONT VIEW

BACK VIEW



RIGHT

LEFT

RIGHT

Version Date: 18.03.10

(6) *In the past 7 days*, how often has your child been tired?

- ☐ Never
- ☐ Almost never
- ☐ Sometimes
- ☐ Often
- ☐ Almost always

There are many possible ways that pain can affect lives of young people. Please respond to each item by checking one box per row.

In the past 7 days…

	Never	Almost Never	Sometimes	Often	Almost Always
(7) It was hard for my child to sleep when he/she had pain.	☐	☐	☐	☐	☐
(8) It was hard for my child to pay attention when he/she had pain.	☐	☐	☐	☐	☐
(9) It was hard for my child to stay standing when he/she had pain.	☐	☐	☐	☐	☐
(10) It was hard for my child to have fun when he/she had pain.	☐	☐	☐	☐	☐
(11) It was hard for my child to do schoolwork when he/she had pain.	☐	☐	☐	☐	☐
(12) It was hard for my child to walk one block when she/he had pain.	☐	☐	☐	☐	☐
(13) It was hard for my child to run when he/she had pain.	☐	☐	☐	☐	☐

Below are some words that describe different feelings and emotions. Read each item and then check the box under the word that describes how often your child has felt this way *in the past 7 days*.

	Never	Almost Never	Sometimes	Often	Almost Always
(14) Sad	☐	☐	☐	☐	☐
(15) Angry	☐	☐	☐	☐	☐
(16) Cheerful	☐	☐	☐	☐	☐
(17) Worried	☐	☐	☐	☐	☐

Thank you for answering these questions.

Child 8-18 years: self-report

SUPER-KIDZ

Recommended Self-Report Version for Children Aged >=8 Years

(1) How much pain do you have *right now*? Check the box below the number that best describes your level of pain right now, with 0 being "no pain" and 10 being "most pain possible."

No pain	0	1	2	3	4	5	6	7	8	9	10	Most pain possible
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

(2) If you had pain *in the past 7 days,* how much did it usually hurt? Check the box below the number that best describes your usual level of pain during *the past 7 days,* with 0 being "no pain" and 10 being "most pain possible."

No pain	0	1	2	3	4	5	6	7	8	9	10	Most pain possible
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

(3) On how many days did you have pain *in the past 7 days?*

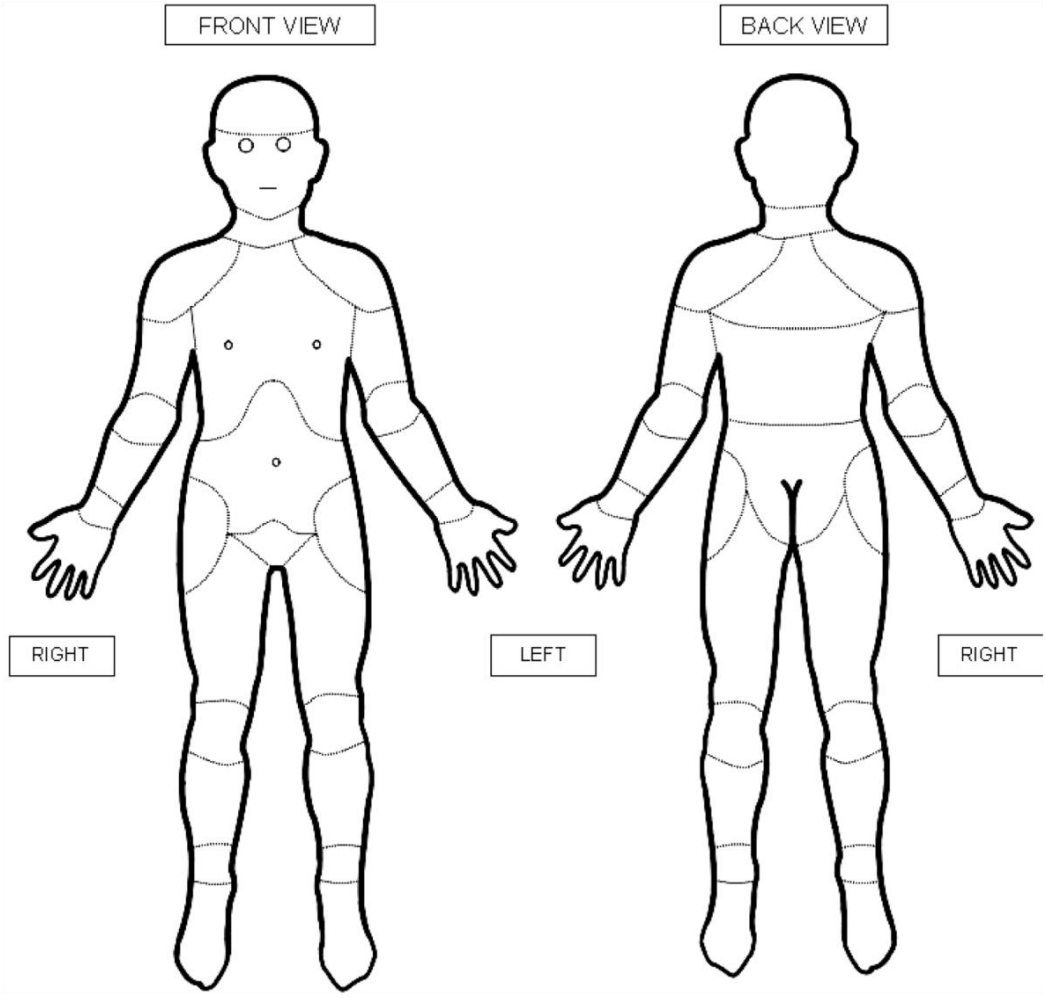
- ☐ 1 day
- ☐ 2 days
- ☐ 3 days
- ☐ 4 days
- ☐ 5 days
- ☐ 6 days
- ☐ Every day

(4) If you had pain *in the past 7 days*, how long did the pain usually last?

- ☐ Less than 30 minutes
- ☐ About 1 hour
- ☐ Between 1 and 3 hours
- ☐ About half the day
- ☐ All day or longer
- ☐ No pain in the past 7 days

Version Date: 18.03.10

(5) Select all the parts of your body where you have had pain *in the past 7 days.*



Version Date: 18.03.10

(6) *In the past 7 days,* how often have you felt tired?

- ☐ Never
- ☐ Almost never
- ☐ Sometimes
- ☐ Often
- ☐ Almost always

There are many possible ways that pain can affect lives of young people. Please respond to each item by checking one box per row.

In the past 7 days…

	Never	Almost Never	Sometimes	Often	Almost Always
(7) I had trouble sleeping when I had pain.	☐	☐	☐	☐	☐
(8) It was hard for me to pay attention when I had pain.	☐	☐	☐	☐	☐
(9) It was hard to stay standing when I had pain.	☐	☐	☐	☐	☐
(10) It was hard to have fun when I had pain.	☐	☐	☐	☐	☐
(11) I had trouble doing schoolwork when I had pain.	☐	☐	☐	☐	☐
(12) It was hard for me to walk one block when I had pain.	☐	☐	☐	☐	☐
(13) It was hard for me to run when I had pain.	☐	☐	☐	☐	☐
(14) I kept thinking about how much I wanted the pain to stop when I had pain.	☐	☐	☐	☐	☐
(15) I was afraid that the pain would get worse when I had pain.	☐	☐	☐	☐	☐
(16) I felt I couldn't stand it anymore when I had pain.	☐	☐	☐	☐	☐

Below are some words that describe different feelings and emotions. Read each item, and then check one box per row under the word that describes how often you felt this way *in the past 7 days*.

	Never	Almost Never	Sometimes	Often	Almost Always
(17) Sad	☐	☐	☐	☐	☐
(18) Angry	☐	☐	☐	☐	☐
(19) Cheerful	☐	☐	☐	☐	☐
(20) Worried	☐	☐	☐	☐	☐

Thank you for answering these questions.

2. Consent and Assent forms

SickKids
THE HOSPITAL FOR SICK CHILDREN
Research Ethics Board

PATIENT CONSENT FORM

Title of Research Project: Validation of **S**tandardized **U**niversal **P**ain **E**valuations for **R**heumatology providers for children and youth (SUPERKIDZ)

Investigator(s):
Principal Investigators:
Dr. Jennifer Stinson, RN, PhD, SickKids, (416) 813-7654 ext. 4514
Co-Investigators:
Dr. Nadia Luca, MD, FRCPC, SickKids, (416) 813-7654 ext. 28030
Dr. Brian Feldman, MD, MSc, FRCPC, SickKids, (416) 813-5828
Dr. Ahmed Bayoumi, MD, MSc, St. Michael's Hospital, (416) 864-5728
Dr. Dorcas Beaton, PhD, St. Michael's Hospital, (416) 864-6060 ext. 77030
Dr. Susanne Benseler, MD, MSc, SickKids, (416) 813-7654 ext. 7711
Dr. Sarah Campillo, MD, Montreal Children's Hospital, (514) 412-4400 ext. 24268
Dr. Claire LeBlanc, MD, Montreal Children's Hospital, (514) 412-4268
Research Nurse Coordinator:
Ms. Navreet Gill, RN, MN, SickKids, (416) 813-7654 ext. 2332
Clinical Research Project Assistant
Margaret van Wyk, BSc (Hons), SickKids, (416) 813-7654 ext. 2314
Summer Research Student
Josip Marcinko, BHSc (Hons), MSc, SickKids, (416) 813-7654 ext.2332

Purpose of the Research:

The goal of this study is to find out whether a new pain tool can correctly measure what type and how much pain pediatric rheumatology patients have.

Description of the Research:

If you agree to take part in this study, you will be asked to complete an online questionnaire (SUPERKIDZ) a few times in the clinic and at home. A research assistant will guide you through the questionnaire on the first occasion. The questionnaire will take about 2-3 minutes. On the initial visit you will be asked to complete 4 additional questionnaires, which will take about 20-30 minutes. There are no extra hospital visits required for this study.

In addition, we will review of your health record to get information on your: age, gender, diagnosis, number of active joints, and your physician's global assessment of how active your disease is. This information will not be identified as being yours.

Potential Harms:

There are no known harms that we are aware of that are associated with participation in this study but there may be harms that we do not know about.

Potential Discomforts or Inconvenience:

This study will involve filling out of multiple questionnaires in the hospital, which will take about 30 minutes. There is also a short amount of time required (2-3 minutes) to complete the questionnaire online from home on two occasions. No extra visits to the hospital are required.

Potential Benefits:

To individual subjects:

You will not benefit directly from participating in this study.

To society:

The standardized pain tool will make it possible to improve the assessment and management of pain for a large number of children with juvenile arthritis across North America, thereby reducing unnecessary suffering and improving health outcomes. This tool will help clinicians to more effectively assess, treat, and monitor pain outcomes in the pediatric rheumatology population. In addition, the computerized measure will allow for the clinicians and patients to use their results immediately during the clinical visit.

It is likely that use of the SUPERKIDZ pain tool will increase patient and physician communication about pain at the clinical encounter, improve patient engagement in decision-making about their pain treatment plan and ultimately lead to improved child health outcomes (such as quality of life).

Confidentiality:

We will respect your privacy. No information about who you are will be given to anyone or be published without your permission, unless required by law. For example, the law could make us give information about you if a child has been abused, if you have an illness that could spread to others, if you or someone else talks about suicide (killing themselves), or if the court orders us to give them the study papers.

SickKids Clinical Research Monitors or the regulator of the study may see your health record to check on the study. By signing this consent form, you agree to let these people look at your records. We will put a copy of this research consent form in your patient health record and give you a copy as well.

The data produced from this study will be stored in a secure, locked location. Only members of the research team (and maybe those individuals described above) will have access to the data. This could include external research team members. Following completion of the research study the data will be kept as long as required then destroyed as required by SickKids policy. Published study results will not reveal your identity.

Reimbursement:

We will reimburse you for all your reasonable out of pocket expenses for being in this study eg., meals, babysitters, parking and getting you to and from SickKids. If you stop taking part in the study, we will pay you for your expenses for taking part in the study up until that point. We will also provide you with some compensation, [$5 gift card], in recognition of your time and effort.

Participation:

It is your choice to take part in this study. You can stop at any time. The care you get at SickKids will not be affected in any way by whether you take part in this study.

New information that we get while we are doing this study may affect your decision to take part in this study. If this happens, we will tell you about this new information. And we will ask you again if you still want to be in the study.

During this study we may create new tests, new medicines, or other things that may be worth some money. Although we may make money from these findings, we cannot give you any of this money now or in the future because you took part in this study.

If you become ill or are harmed because of study participation, we will treat you for free. Your signing this consent form does not interfere with your legal rights in any way. The staff of the study, any people who gave money for the study, or the hospital are still responsible, legally and professionally, for what they do.

Sponsorship:

The funders of this research are Dr. Jennifer Stinson and The Hospital for Sick Children.

Conflict of Interest:

The Principal Investigator, Dr. Jennifer Stinson, and the other research team members have no conflict of interest to declare.

Consent:

By signing this form, I agree that:

1) You have explained this study to me. You have answered all my questions.

2) You have explained the possible harms and benefits (if any) of this study.

3) I know what I could do instead of taking part in this study. I understand that I have the right not to take part in the study and the right to stop at any time. My decision about taking part in the study will not affect my health care at SickKids.

4) I am free now, and in the future, to ask questions about the study.

5) I have been told that my medical records will be kept private except as described to me.

6) I understand that no information about who I am will be given to anyone or be published without first asking my permission.

7) I agree, or consent, to take part in this study.

_____	_____
Printed Name of Subject & Age	Subject's signature & date
_____	_____
Printed Name of person who explained consent	Signature of Person who explained consent & date
_____	_____
Printed Witness' name (if the subject/legal guardian does not read English)	Witness' signature & date

If you have any questions about this study, please call Dr. Jennifer Stinson at (416) 813-7654 ext. 4514.

If you have questions about your rights as a subject in a study or injuries during a study, please call the Research Ethics Manager at 416-813-5718.

SickKids
THE HOSPITAL FOR SICK CHILDREN
Research Ethics Board

PARENT CONSENT FORM

Title of Research Project: Validation of **S**tandardized **U**niversal **P**ain **E**valuations for **R**heumatology providers for children and youth (SUPERKIDZ)

Investigator(s):

Principal Investigators:
Dr. Jennifer Stinson, RN, PhD, SickKids, (416) 813-7654 ext. 4514
Co-Investigators:
Dr. Nadia Luca, MD, FRCPC, SickKids, (416) 813-7654 ext. 28030
Dr. Brian Feldman, MD, MSc, FRCPC, SickKids, (416) 813-5828
Dr. Ahmed Bayoumi, MD, MSc, St. Michael's Hospital, (416) 864-5728
Dr. Dorcas Beaton, PhD, St. Michael's Hospital, (416) 864-6060 ext. 77030
Dr. Susanne Benseler, MD, MSc, SickKids, (416) 813-7654 ext. 7711
Dr. Sarah Campillo, MD, Montreal Children's Hospital, (514) 412-4400 ext. 24268
Dr. Claire LeBlanc, MD, Montreal Children's Hospital, (514) 412-4268
Research Nurse Coordinator:
Ms. Navreet Gill, RN, MN, SickKids, (416) 813-7654 ext. 2332
Clinical Research Project Assistant
Margaret van Wyk, BSc (Hons), SickKids, (416) 813-7654 ext. 2314
Summer Research Student
Josip Marcinko, BHSc (Hons), MSc, SickKids, (416) 813-7654 ext.2332

Purpose of the Research:

The goal of this study is to find out whether a new pain tool can correctly measure what type and how much pain pediatric rheumatology patients have.

Description of the Research:

If you and your child agree to take part in this study, you and/or your child will be asked to complete an online questionnaire (SUPERKIDZ) a few times in the clinic and at home. A research assistant will guide you through the questionnaire on the first occasion. The questionnaire will take about 2-3 minutes. On the initial visit you will be asked to complete 4 additional questionnaires, which will take about 20-30 minutes. There are no extra hospital visits required for this study.

In addition, we will review of your child's health record for the research project to get information on: age, gender, diagnosis, number of active joints, and your physician's global assessment of how active your disease is. This information will not be identified as being your child's.

Potential Harms:

There are no known harms that we are aware of that are associated with participation in this study but there may be harms that we do not know about.

Potential Discomforts or Inconvenience:

This study will involve filling out multiple questionnaires in the hospital, which will take about 30 minutes. There is also a short amount of time required (2-3 minutes) to complete the questionnaire online from home on two occasions. No extra visits to the hospital are required.

Potential Benefits:

To individual subjects:

You will not benefit directly from participating in this study.

To society:

The standardized pain tool will make it possible to improve the assessment and management of pain for a large number of children with juvenile arthritis across North America, thereby reducing unnecessary suffering and improving health outcomes. This tool will help clinicians to more effectively assess, treat, and monitor pain outcomes in the pediatric rheumatology population. In addition, the computerized measure will allow for the clinicians and patients to use their results immediately during the clinical visit.

It is likely that use of the SUPERKIDZ pain tool will increase patient and physician communication about pain at the clinical encounter, improve patient engagement in decision-making about their pain treatment plan and ultimately lead to improved child health outcomes (such as quality of life).

Confidentiality:

We will respect your privacy. No information about who you are will be given to anyone or be published without your permission, unless required by law. For example, the law could make us give information about you if a child has been abused, if you have an illness that could spread to others, if you or someone else talks about suicide (killing themselves), or if the court orders us to give them the study papers.

SickKids Clinical Research Monitors or the regulator of the study may see your health record to check on the study. By signing this consent form, you agree to let these people look at your records. We will put a copy of this research consent form in your patient health record and give you a copy as well.

The data produced from this study will be stored in a secure, locked location. Only members of the research team (and maybe those individuals described above) will have access to the data. This could include external research team members. Following completion of the research study the data will be kept as long as required then destroyed as required by SickKids policy. Published study results will not reveal your identity.

Reimbursement:

We will reimburse you for all your reasonable out of pocket expenses for being in this study eg., meals, babysitters, parking and getting you to and from SickKids. If you stop taking part in the study, we will pay you for your expenses for taking part in the study up until that point. We will also provide you with some compensation, [$5 gift card], in recognition of your time and effort.

Participation:

It is your choice to take part in this study. You can stop at any time. The care you get at SickKids will not be affected in any way by whether you take part in this study.

New information that we get while we are doing this study may affect your decision to take part in this study. If this happens, we will tell you about this new information. And we will ask you again if you still want to be in the study.

During this study we may create new tests, new medicines, or other things that may be worth some money. Although we may make money from these findings, we cannot give you any of this money now or in the future because you took part in this study.

If you become ill or are harmed because of study participation, we will treat you for free. Your signing this consent form does not interfere with your legal rights in any way. The staff of the study, any people who gave money for the study, or the hospital are still responsible, legally and professionally, for what they do.

Sponsorship:

The funders of this research are Dr. Jennifer Stinson and The Hospital for Sick Children.

Conflict of Interest:

The Principal Investigator, Dr. Jennifer Stinson, and the other research team members have no conflict of interest to declare.

Consent:

By signing this form, I agree that:

1) You have explained this study to me. You have answered all my questions.
2) You have explained the possible harms and benefits (if any) of this study.
3) I know what I could do instead of taking part in this study. I understand that I have the right not to take part in the study and the right to stop at any time. My decision about taking part in the study will not affect my health care at SickKids.
4) I am free now, and in the future, to ask questions about the study.
5) I have been told that my medical records will be kept private except as described to me.
6) I understand that no information about who I am will be given to anyone or be published without first asking my permission.
7) I agree, or consent, to take part in this study.

_____	_____
Printed Name of Parent	Parent's signature & date
_____	_____
Printed Name of person who explained consent	Signature of Person who explained consent & date
_____	_____
Printed Witness' name (if the subject/legal guardian does not read English)	Witness' signature & date

If you have any questions about this study, please call Dr. Jennifer Stinson at (416) 813-7654 ext. 4514.

If you have questions about your rights as a subject in a study or injuries during a study, please call the Research Ethics Manager at 416-813-5718.

SickKids
THE HOSPITAL FOR SICK CHILDREN
Research Ethics Board

ASSENT FORM

Title of Study: Validation of **S**tandardized **U**niversal **P**ain **E**valuations for **R**heumatology providers for children and youth (SUPERKIDZ)

Investigator(s):

Principal Investigators:
Dr. Jennifer Stinson, RN, PhD, SickKids, (416) 813-7654 ext. 4514
Co-Investigators:
Dr. Nadia Luca, MD, FRCPC, SickKids, (416) 813-7654 ext. 28030
Dr. Brian Feldman, MD, MSc, FRCPC, SickKids, (416) 813-5828
Dr. Ahmed Bayoumi, MD, MSc, St. Michael's Hospital, (416) 864-5728
Dr. Dorcas Beaton, PhD, St. Michael's Hospital, (416) 864-6060 ext. 77030
Dr. Susanne Benseler, MD, MSc, SickKids, (416) 813-7654 ext. 7711
Dr. Sarah Campillo, MD, Montreal Children's Hospital, (514) 412-4400 ext. 24268
Dr. Claire LeBlanc, MD, Montreal Children's Hospital, (514) 412-4268
Research Nurse Coordinator:
Ms. Navreet Gill, RN, MN, SickKids, (416) 813-7654 ext. 2332
Clinical Research Project Assistant
Margaret van Wyk, BSc (Hons), SickKids, (416) 813-7654 ext. 2314
Summer Research Student
Josip Marcinko, BHSc (Hons), MSc, SickKids, (416) 813-7654 ext.2332

Why are we doing this study?

To test if the new SUPERKIDZ questionnaire can measure your real pain.

What will happen during the study?

This SUPERKIDZ questionnaire is a program on the computer that helps us measure different things about pain, like how much pain you are in and where the pain might be. During the study we will show you how to use it and then we will ask you to complete it on a few different days. You and your parents will also fill out a few more questionnaires during your clinic visit. We will also get some information about your arthritis from your hospital file.

Are there good things and bad things about the study?

There are no bad things about the study that we can think of. The good thing about the study is that it gives you a chance to talk about your pain.

Who will know about what I did in the study?

If we feel your health may be in danger, we may have to report your results to your doctor.

Can I decide if I want to be in the study?

Nobody will be angry or upset if you do not want to be in the study. We are talking to your parent/legal guardians about the study and you should talk to them about it too.

Assent:

I was present when _____ read this form and said that he or she agreed, or assented, to take part in this study.

_____ _____

Printed Name of person who obtained assent Signature & Date

3. Patient health information form

Unique Study ID: _____

Meds for JIA		Meds for Other Conditions	
Medication Name	**Dose**	**Medication Name**	**Dose**
❍ Naproxen (Naprosyn)		❍ Other (name: _____)	
❍ Ibuprofen (Motrin, Advil)		❍ Other (name: _____)	
❍ Indomethacin (Indocid)		❍ Other (name: _____)	
❍ Acetominophen (Tylenol)		❍ Other (name: _____)	
❍ Oxycodone		❍ Other (name: _____)	
❍ Methadone			
❍ Methotrexate			
❍ Leflunomide (Arava)			
❍ Sulfasalazine			
❍ Etanercept (Enbrel)			
❍ Infliximab (Remicade)			
❍ Adalimumab (Humira)			
❍ Tocilizumab (Cimzia)			
❍ Lenalidomide (Revlimid)			
❍ Abatacept (Orencia)			
❍ Anakinra (Kineret)			
❍ Rituximab (Rituxan)			
❍ Calcium supplement			
❍ Vitamin D supplement			

1. JIA subtype:

□ Persistent Oligo JIA
□ Extended Oligo JIA
□ Poly JIA (RF negative)
□ Poly JIA (RF positive)
□ Enthesitis-related arthritis
□ Psoriatic arthritis
□ Systemic-onset JIA
□ Undifferentiated

2. Physician's Global Assessment of Disease Activity Rating (Obtain from Rheumatologist):

No activity □0 □1 □2 □3 □4 □5 □6 □7 □8 □9 □10 Maximum activity

3. Total number of active joints: _____

4. Summary of study procedures and timeline by recruitment group

	Group A		**Group B**
Parameter	Test-retest reliability Responsiveness		Test-retest reliability
Sample size	50 per age group (100 total)		30
Inclusion criteria	a) 4-18 years b) fluency in English c) diagnosed with JIA by a rheumatologist d) have access to a computer <u>Additional:</u> e) scheduled to undergo JI		a) 4-18 years b) fluency in English c) diagnosed with JIA by a rheumatologist d) have access to a computer <u>Additional:</u> e) no anticipated changes to management
Initial clinic visit	SUPER-KIDZ*	**Initial clinic visit**	SUPER-KIDZ*
Day before joint injection (JI)	5-Point GRCP SUPER-KIDZ*#	**1 week later**	5-Point GRCP SUPER-KIDZ*
1 week after JI	5-Point GRCP SUPER-KIDZ#		
2 weeks after JI	5-Point GRCP SUPER-KIDZ*#		
3 weeks after JI	5-Point GRCP SUPER-KIDZ*		

* For test-retest reliability

For responsiveness

5-Point GRCP=Five-Point Global Rating of Change in Pain

5. Five-point global rating of change in pain (GRCP)

Think about your pain over the last seven days and compare it to the pain you have today. How has it CHANGED?

Please tick only one box.

☐	☐	☐	☐	☐
Much worse	A little worse	The same	A little better	Much better

6. Data dictionary of variables

Variable	Code/Value	Data type
Gender	M=1 F=0	Categorical
Age (years)	4.0-18.0	Continuous
poligo = persistent oligo	Y=1 N=0	Categorical
poligo = extended oligo	Y=1 N=0	Categorical
polyRF_= poly RF negative	Y=1 N=0	Categorical
polyRF+ = poly RF positive	Y=1 N=0	Categorical
Psoriatic	Y=1 N=0	Categorical
Systemic	Y=1 N=0	Categorical
Undiff = undifferentiated	Y=1 N=0	Categorical
num_meds	0-8	Ordinal
NSAID = ibuprofen, naproxen, indomethacin, diclofenac	Y=1 N=0	Categorical
steroid = prednisone, methylprednisolone	Y=1 N=0	Categorical
DMARD = methtorexate, leflunomide, sulfasalazine	Y=1 N=0	Categorical
anti_TNF = etanercept, infliximab, adalimumab, golimumab	Y=1 N=0	Categorical
other_biol = abatacept, tocilizumab, anakinra, rituximab	Y=1 N=0	Categorical
tylenol	Y=1 N=0	Categorical
narcotic = methadone, oxycodone, morphine	Y=1 N=0	Categorical
Ca_vitD =calcium or vitamin D	Y=1 N=0	Categorical
active_jts = number of active joints	0-unlimited	Interval
PGA = physician global assessment on 10 cm VAS	0-10	Continuous
Q1 = current pain intensity	0-10	Continuous
Q2 = average pain intensity over past week	0-10	Continuous
Q3 = how many days had pain in the past week	1-7	Ordinal
Q4 = how long did pain usually last	1-6	Ordinal
Q5 = all parts of body where have pain	1-59	Continuous
Q6 = how often felt tired in past week	0-4	Ordinal
Q7 = trouble sleeping when in pain	0-4	Ordinal
Q8 = hard to pay attention when in pain	0-4	Ordinal
Q9 = hard to stay standing when in pain	0-4	Ordinal
Q10 = hard to have fun when in pain	0-4	Ordinal
Q11 = trouble doing schoolwork when in pain	0-4	Ordinal
Q12 = hard to walk one block when in pain	0-4	Ordinal
Q13 = hard to run when in pain	0-4	Ordinal
Q14 = keep thinking about how much wanted pain to stop	0-4	Ordinal
Q15 = afraid pain would get worse	0-4	Ordinal
Q16 = could not stand it anymore when in pain	0-4	Ordinal
Q17 = how often feel sad	0-4	Ordinal
Q18 = how often feel angry	0-4	Ordinal
Q19 = how often feel cheerful (NB reverse ordered)	0-4	Ordinal
Q20 = how often feel worried	0-4	Ordinal
GRCP = global rating of change in pain over the previous 1 week. Much worse =1, a little worse =2, same =3, a little better =4, much better =5	1-5	Categorical

7. Distribution of SUPER-KIDZ items at baseline (table and figures)

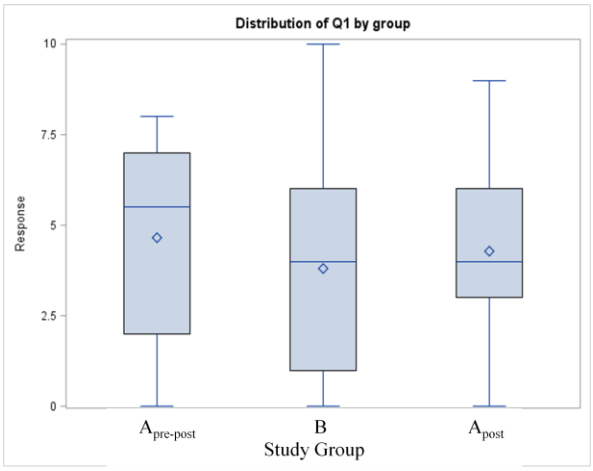
Table: Summary of baseline SUPER-KIDZ data by item (n=51)

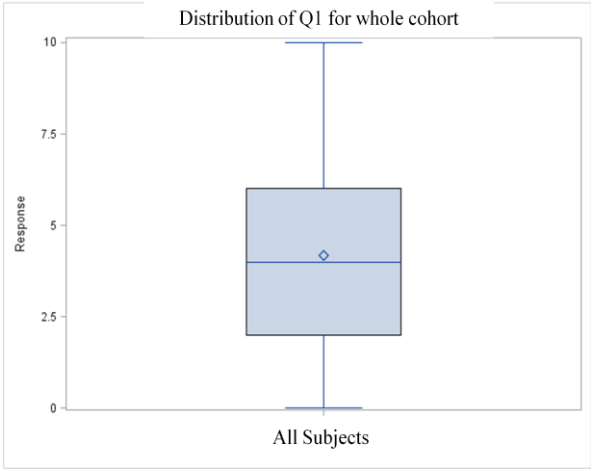
Domain	Item #	# Missing	Response											Item mean	Item median	SD	IQR
			0	1	2	3	4	5	6	7	8	9	10				
Sensory	1	0	8	4	3	4	7	7	8	3	5	1	1	4.1	4.0	2.8	2.0-6.0
	2	0	3	2	6	4	8	7	6	9	4	2	0	4.7	5.0	2.4	3.0-7.0
	3	0	-	5	3	6	10	7	3	17	-	-	-	4.7	5.0	2.0	3.0-7.0
	4	0	-	1	9	5	9	10	17	-	-	-	-	4.4	5.0	1.6	3.0-6.0
	5	0	-	-	-	-	-	-	-	-	-	-	-	7.5	5.0	8.4	2.0-9.0
Fatigue	6	0	4	1	23	16	7	-	-	-	-	-	-	2.4	2.0	1.0	2.0-3.0
Inter-ference	7	0	21	9	17	2	2	-	-	-	-	-	-	1.1	1.0	1.1	0-2.0
	8	0	16	13	12	9	1	-	-	-	-	-	-	1.3	1.0	1.2	0-2.0
	9	0	11	10	14	9	7	-	-	-	-	-	-	1.8	2.0	1.3	1.0-3.0
	10	0	11	12	11	12	5	-	-	-	-	-	-	1.8	2.0	1.3	1.0-3.0
	11	0	20	6	15	8	2	-	-	-	-	-	-	1.3	1.0	1.3	0-2.0
	12	0	11	8	20	7	5	-	-	-	-	-	-	1.8	2.0	1.2	1.0-2.0
	13	0	6	4	18	9	14	-	-	-	-	-	-	2.4	2.0	1.3	2.0-4.0
	14	0	6	5	10	17	13	-	-	-	-	-	-	2.5	3.0	1.3	2.0-4.0
	15	0	9	5	22	11	4	-	-	-	-	-	-	1.9	2.0	1.2	1.0-3.0
	16	0	13	14	12	7	5	-	-	-	-	-	-	1.5	1.0	1.3	0-2.0
Emotional	17	0	15	13	13	10	0	-	-	-	-	-	-	1.4	1.0	1.1	0-2.0
	18	0	19	13	14	5	0	-	-	-	-	-	-	1.1	1.0	1.0	0-2.0
	19	0	7	12	24	4	4	-	-	-	-	-	-	1.7	2.0	1.1	1.0-2.0
	20	0	13	8	18	8	4	-	-	-	-	-	-	1.6	2.0	1.2	0-2.0

SD=standard deviation; IQR=interquartile range

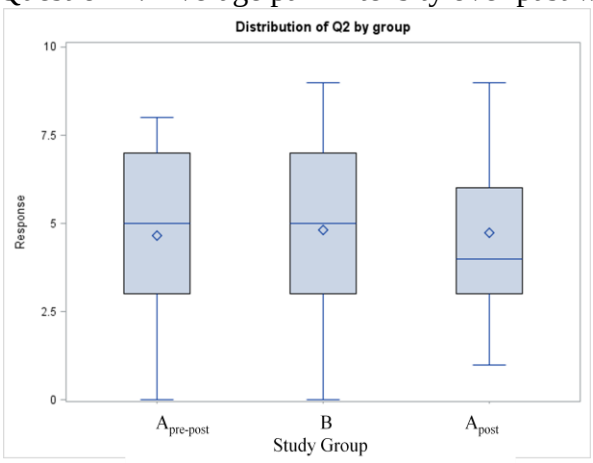
Figures: Boxplots of baseline item scores (n=51). Diamond = mean score, Horizontal blue line = median score, Box = 25% and 75% quartiles, Whiskers = range.

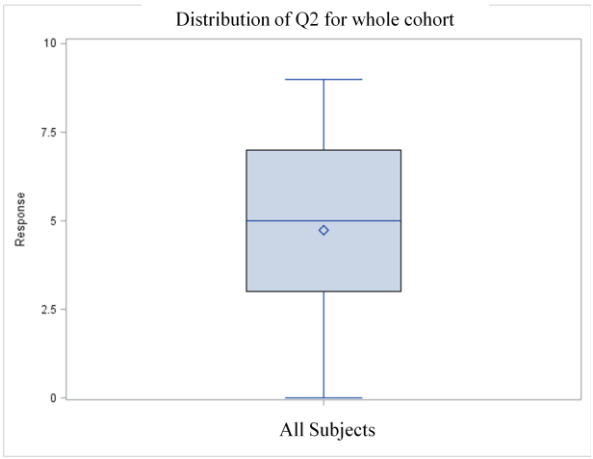
Question 1: Current pain intensity





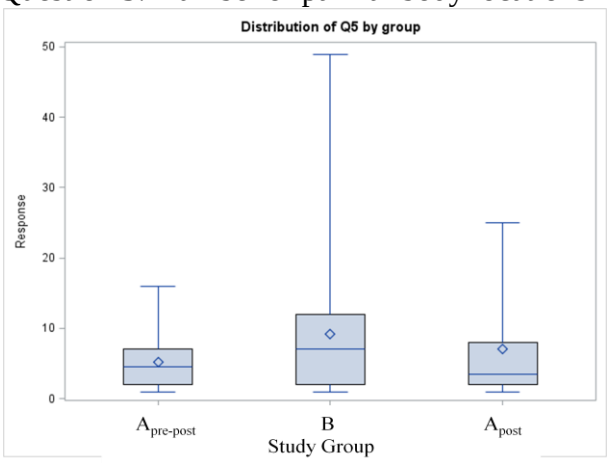
Question 2: Average pain intensity over past week

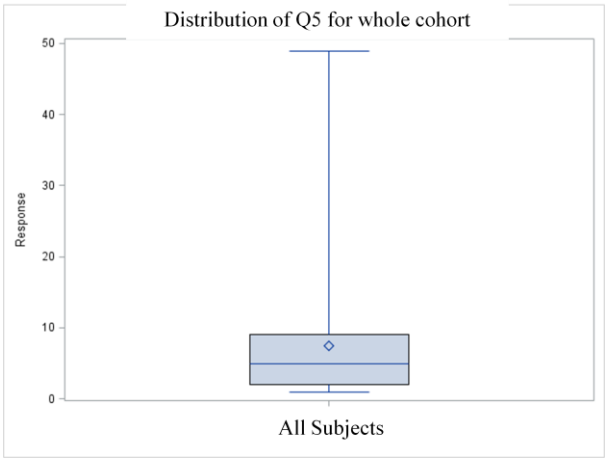




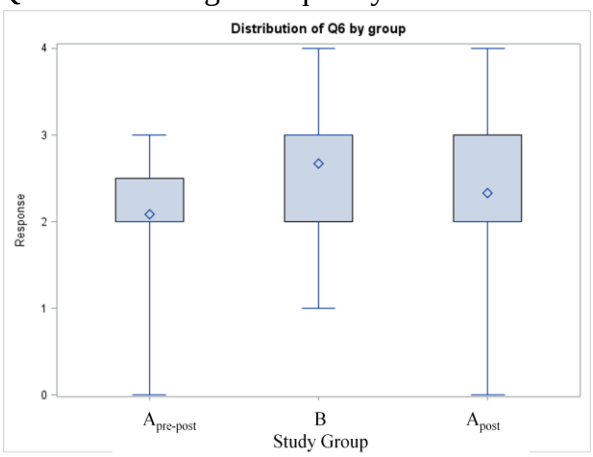
Questions 3 and 4: in Results section

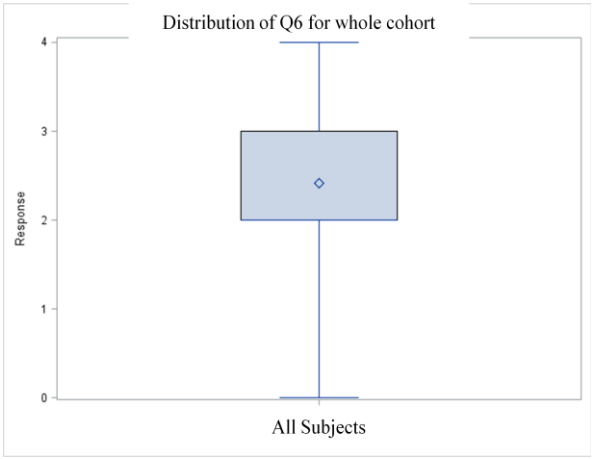
Question 5: Number of painful body locations





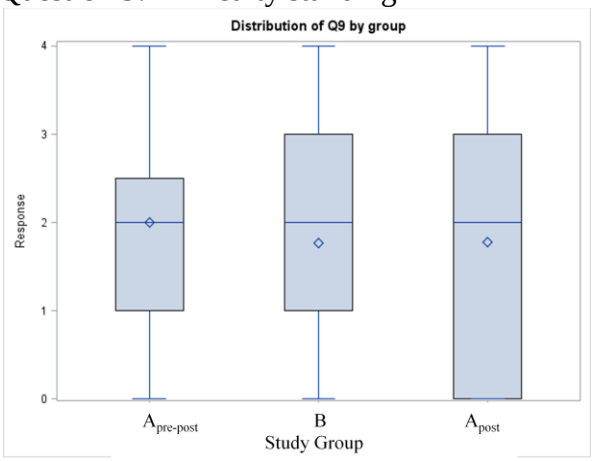
Question 6: Fatigue frequency

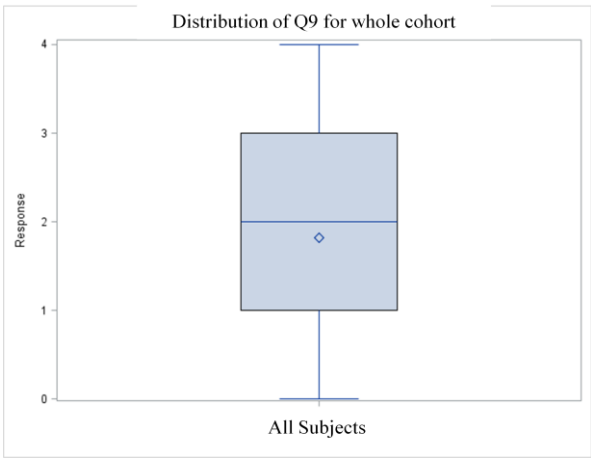




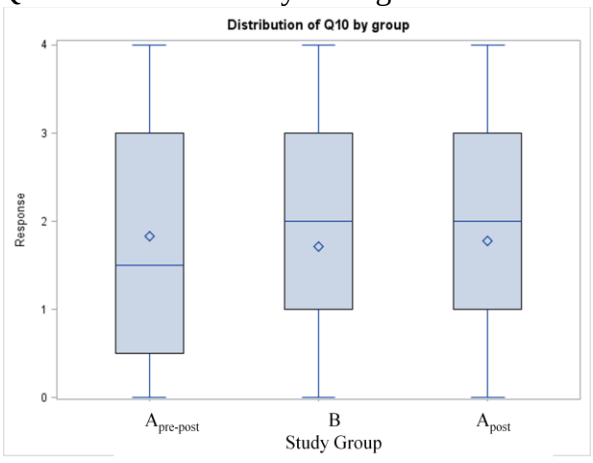
Questions 7 & 8: in Results section

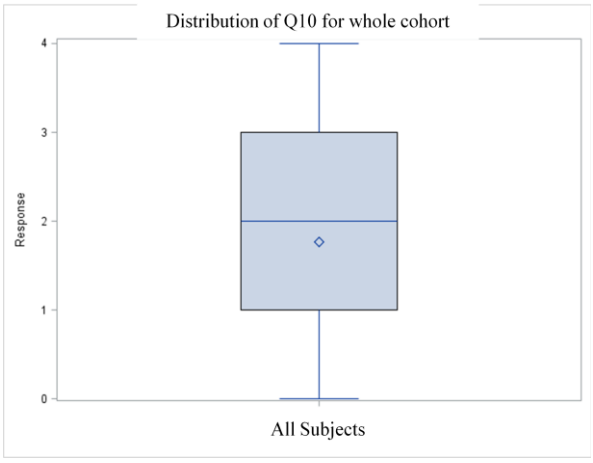
Question 9: Difficulty standing





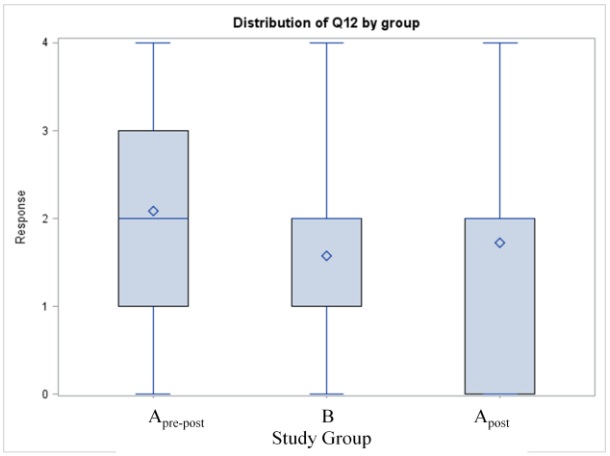
Question 10: Difficulty having fun

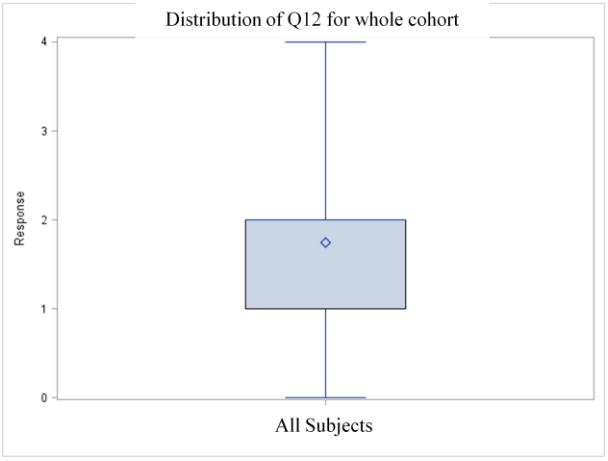




Question 11: In Results section

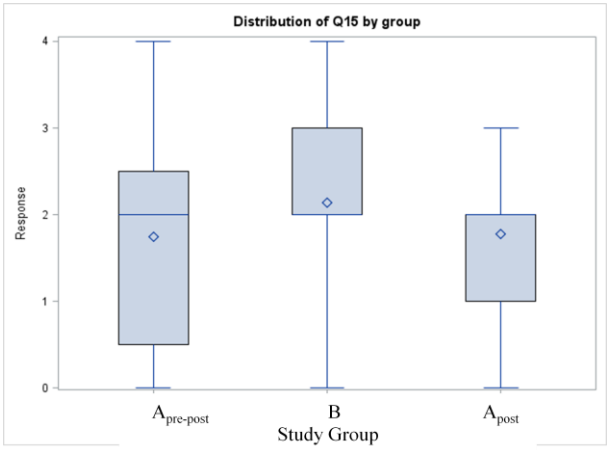
Question 12: Difficulty walking one block

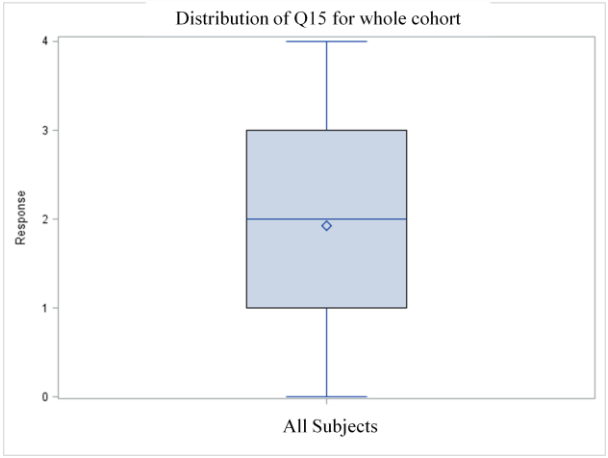




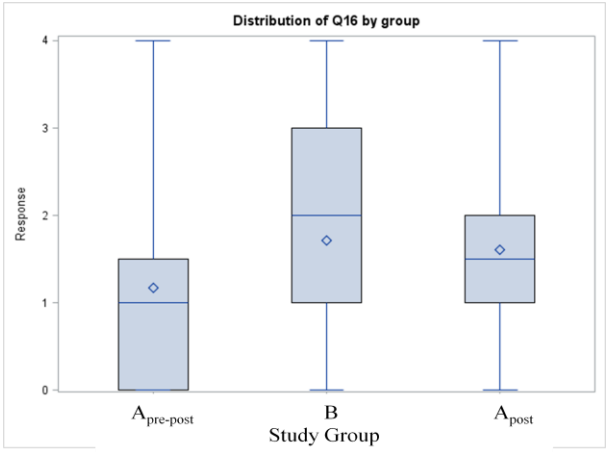
Questions 13 & 14: in Results section

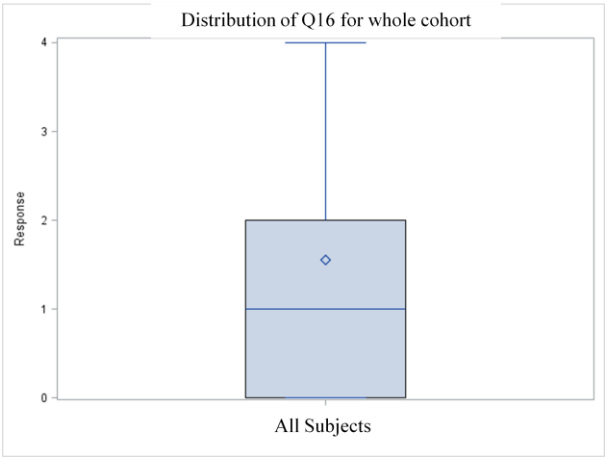
Question 15: Afraid pain would get worse





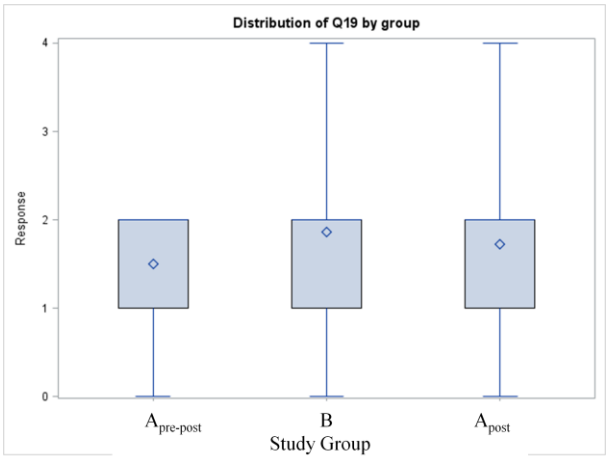
Question 16: Couldn't stand it anymore

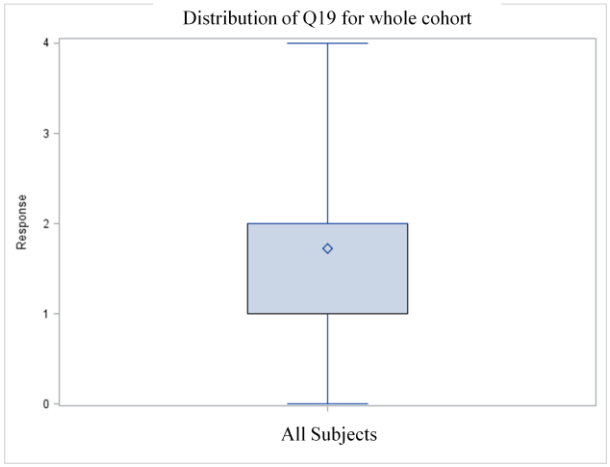




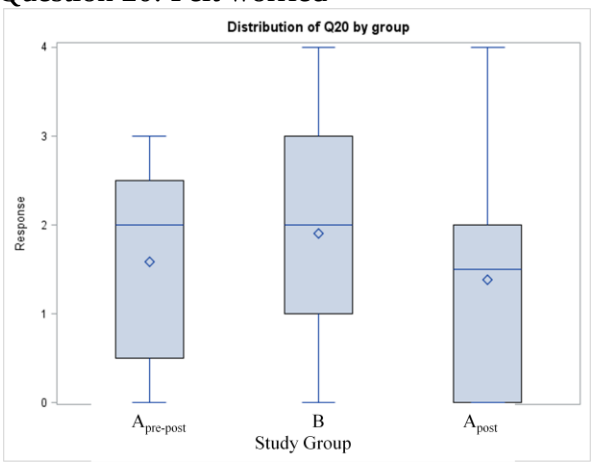
Questions 17 & 18: in Results section

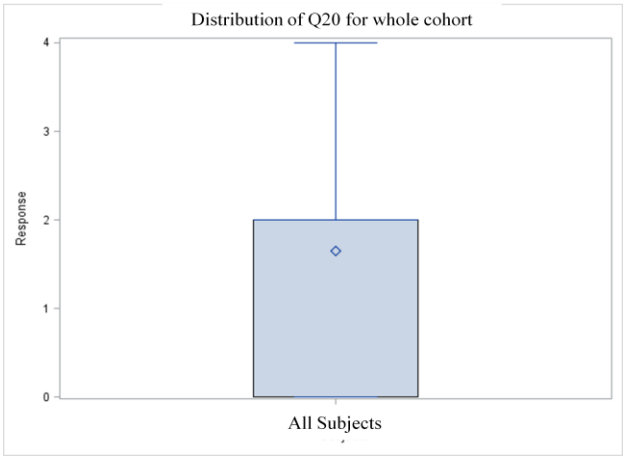
Question 19: Felt cheerful





Question 20: Felt worried



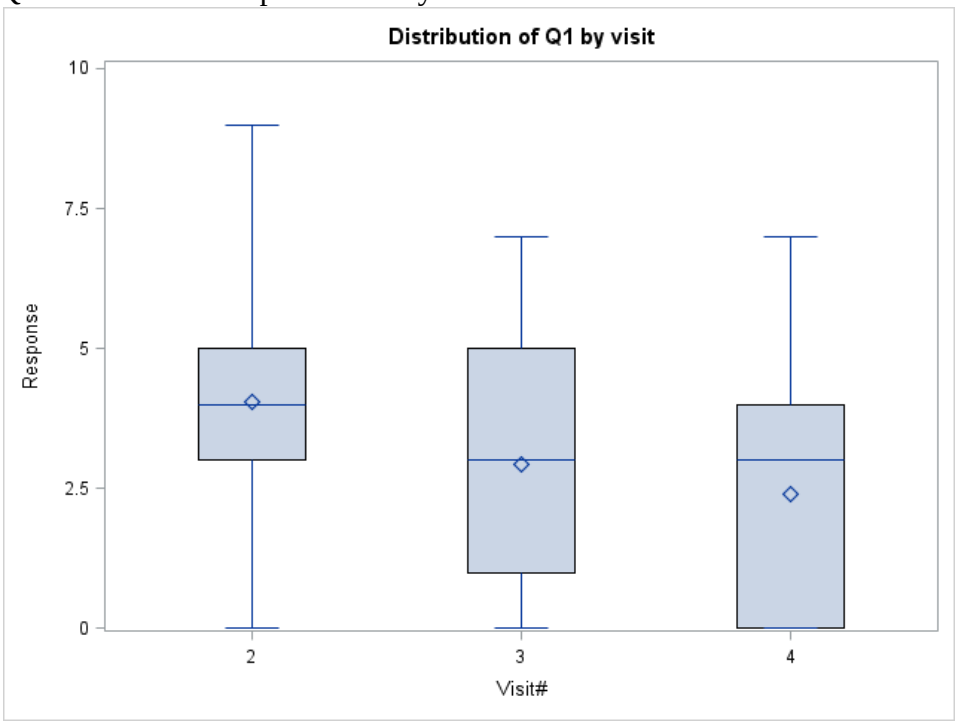


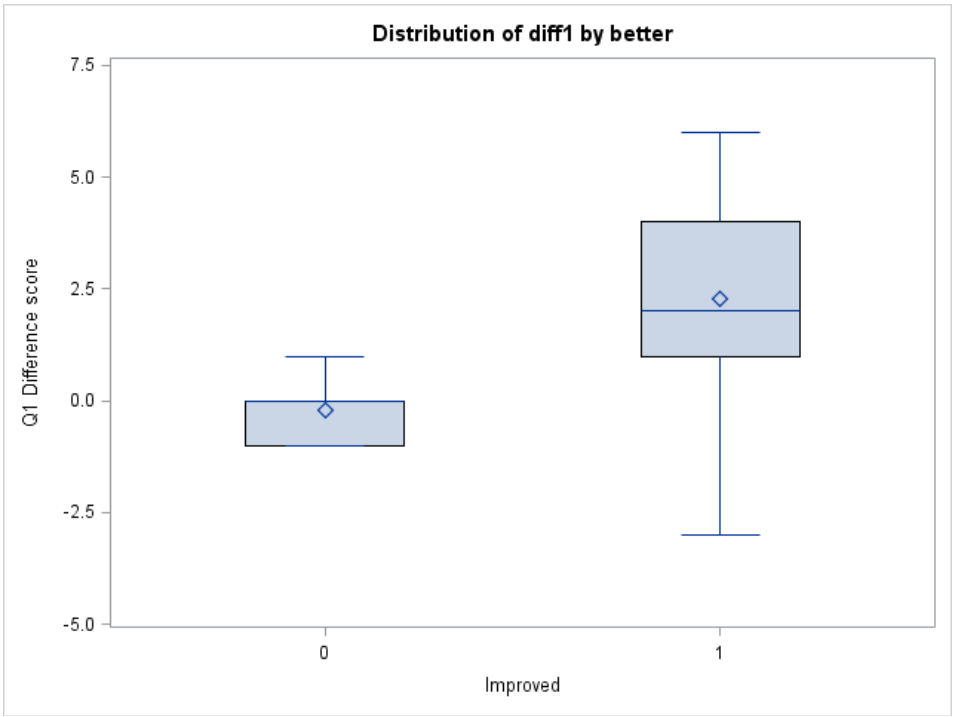
8. Distribution of SUPER-KIDZ item scores over time (figures and table)

<u>Figure:</u> Boxplots of scores over time and plot of change score in improved versus not improved subjects at 2 weeks post-injection for each SUPER-KIDZ item.

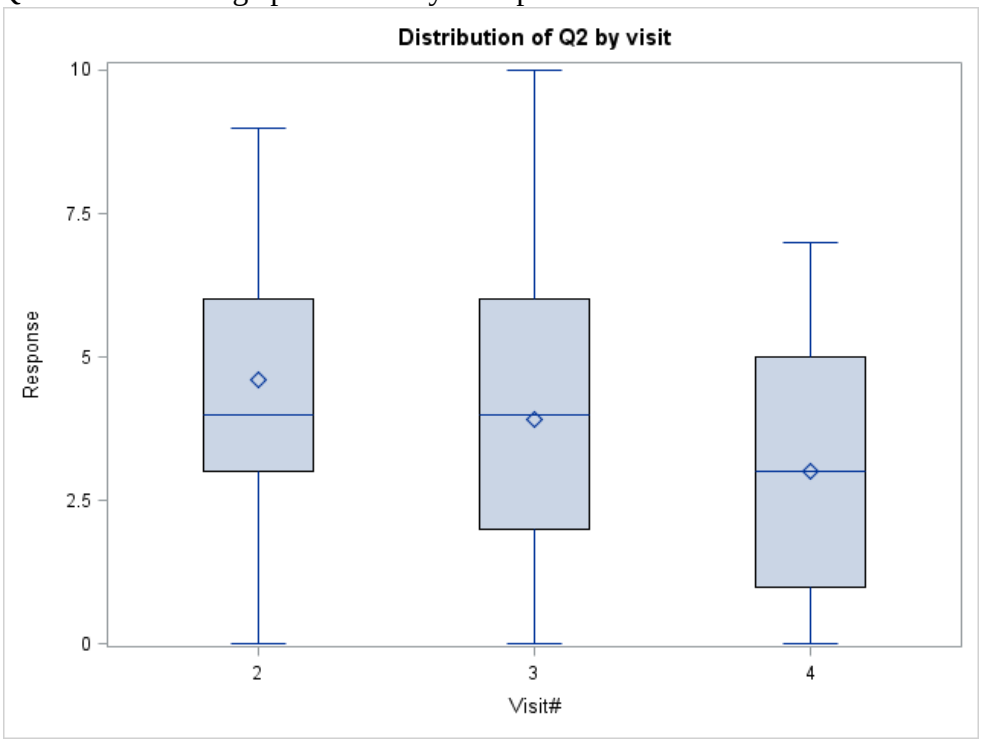
Diamond = mean, Horizontal line = median, Box = 25% and 75% quartiles, Whiskers = range. Visit#: 2=pre-joint injection, 3=1 week post-injection, 4=2 weeks post-injection. Improved: 1=yes 0=no.

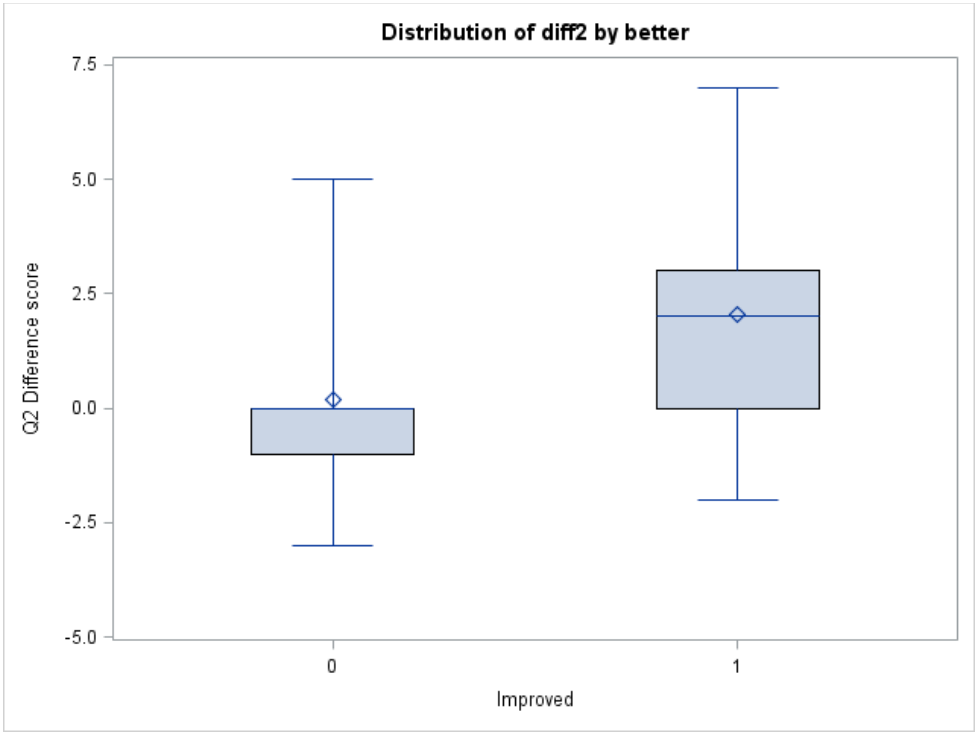
Question 1: Current pain intensity



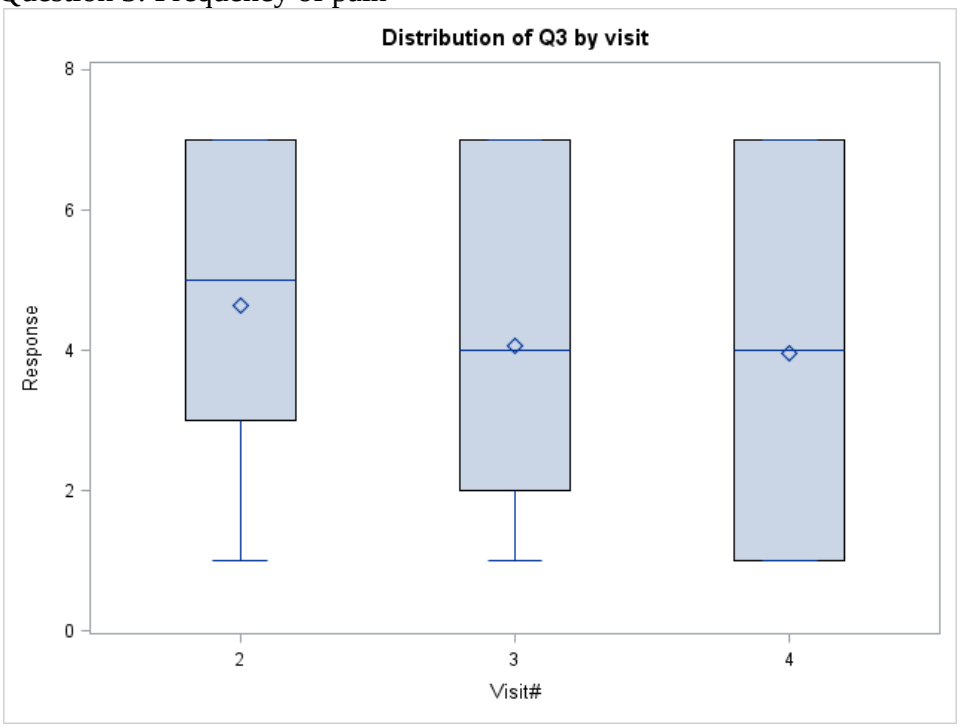


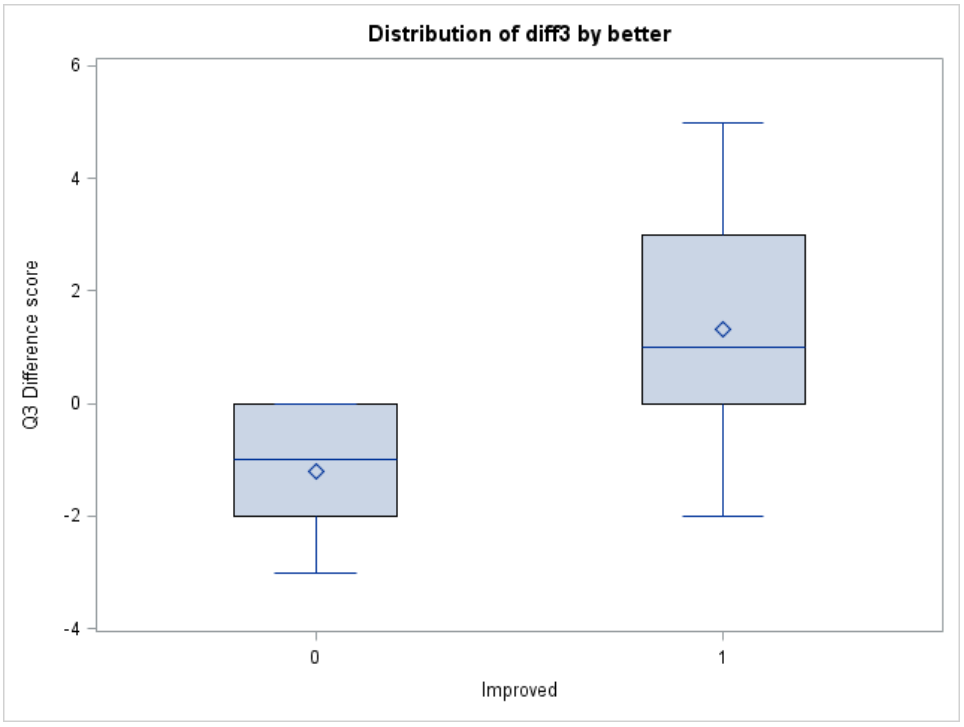
Question 2: Average pain intensity over past week



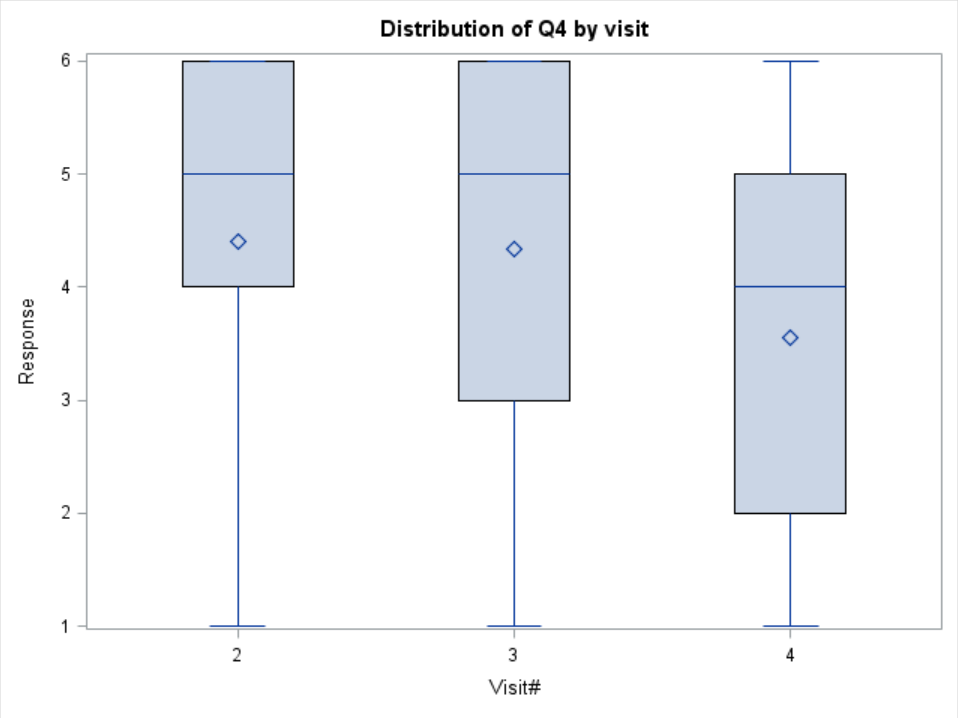


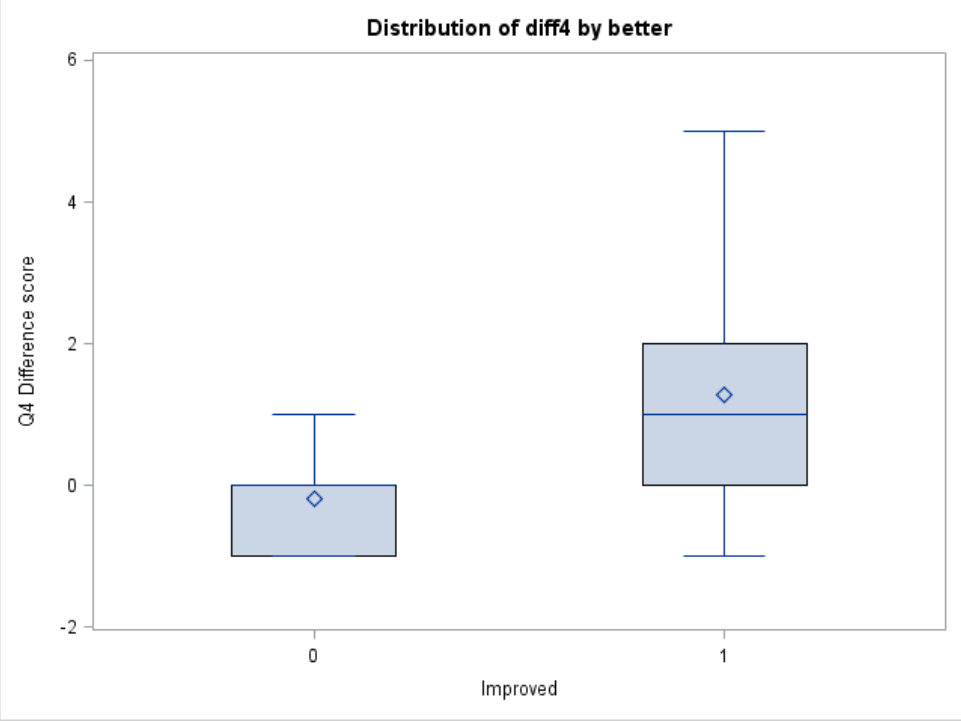
Question 3: Frequency of pain



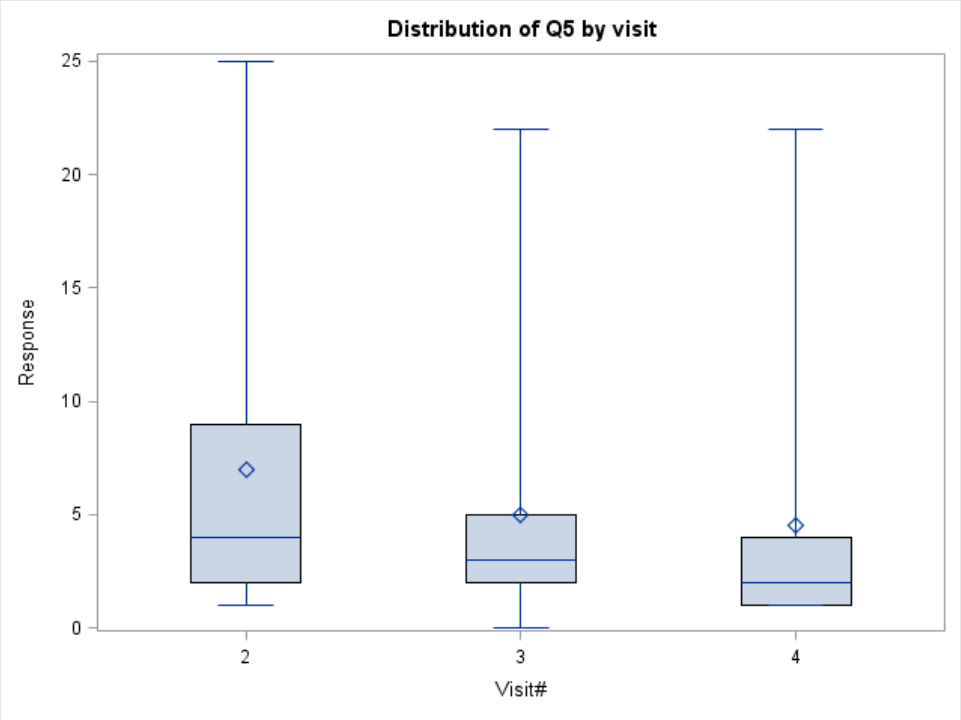


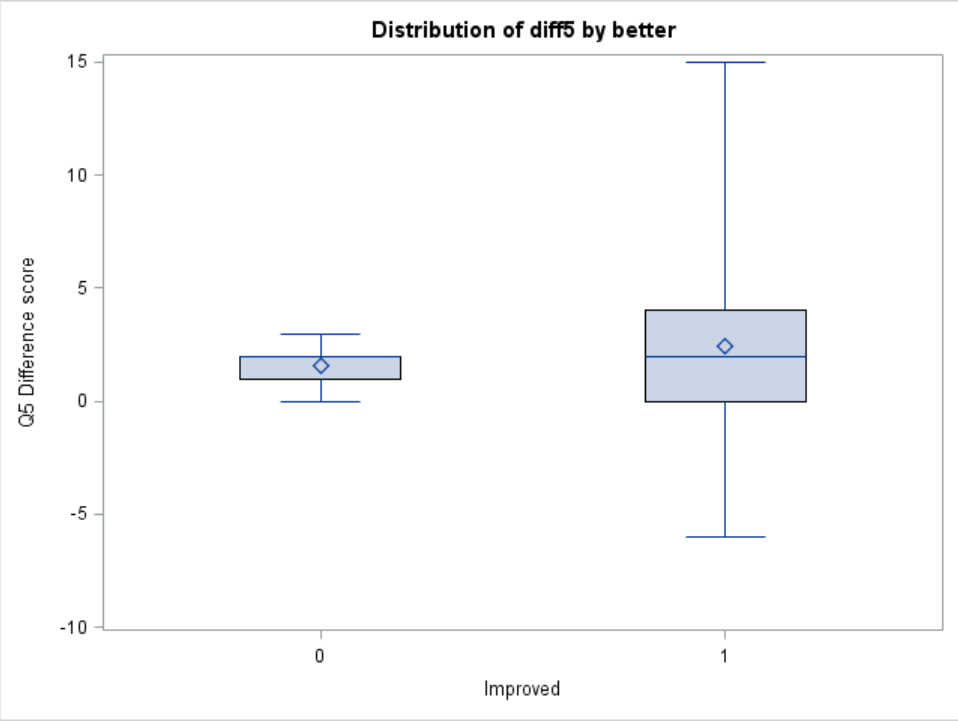
Question 4: Duration of pain



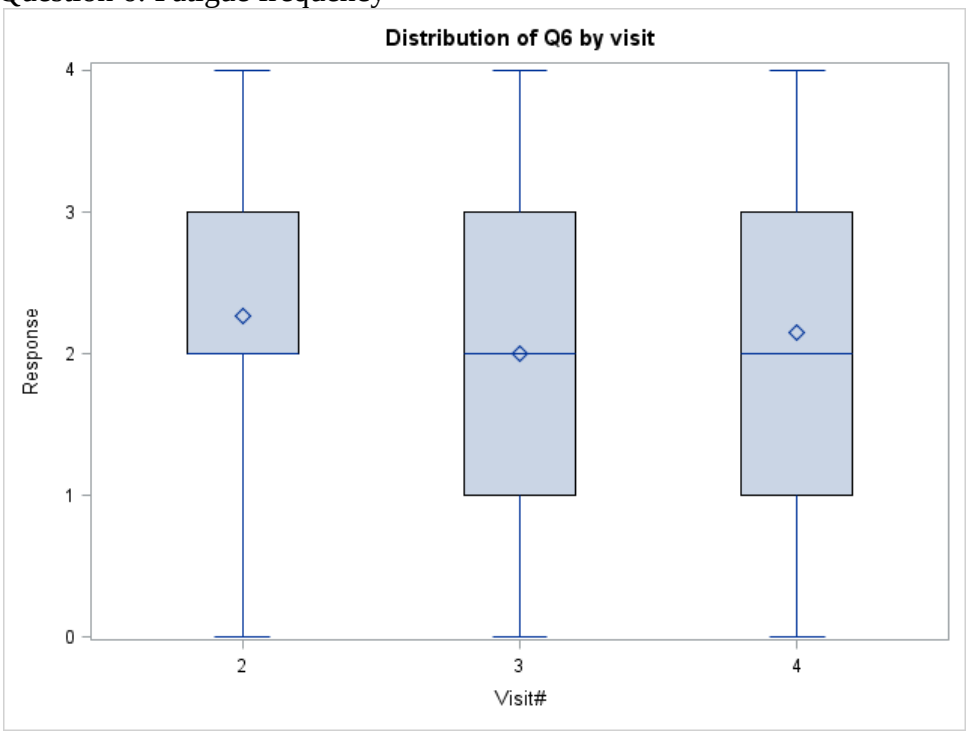


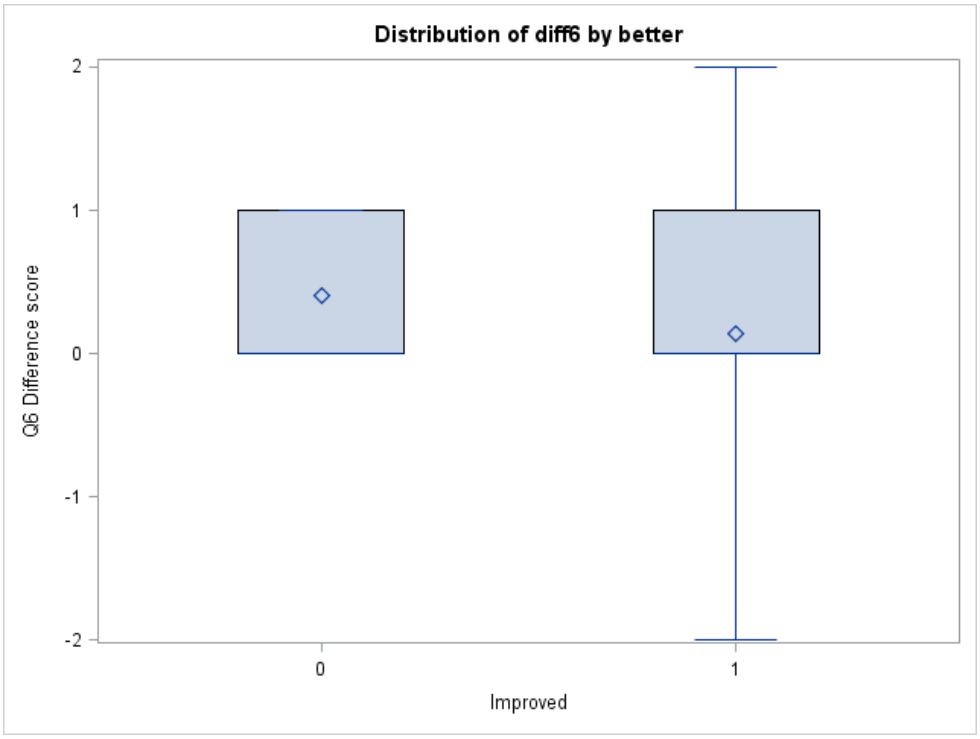
Question 5: Number of painful body locations



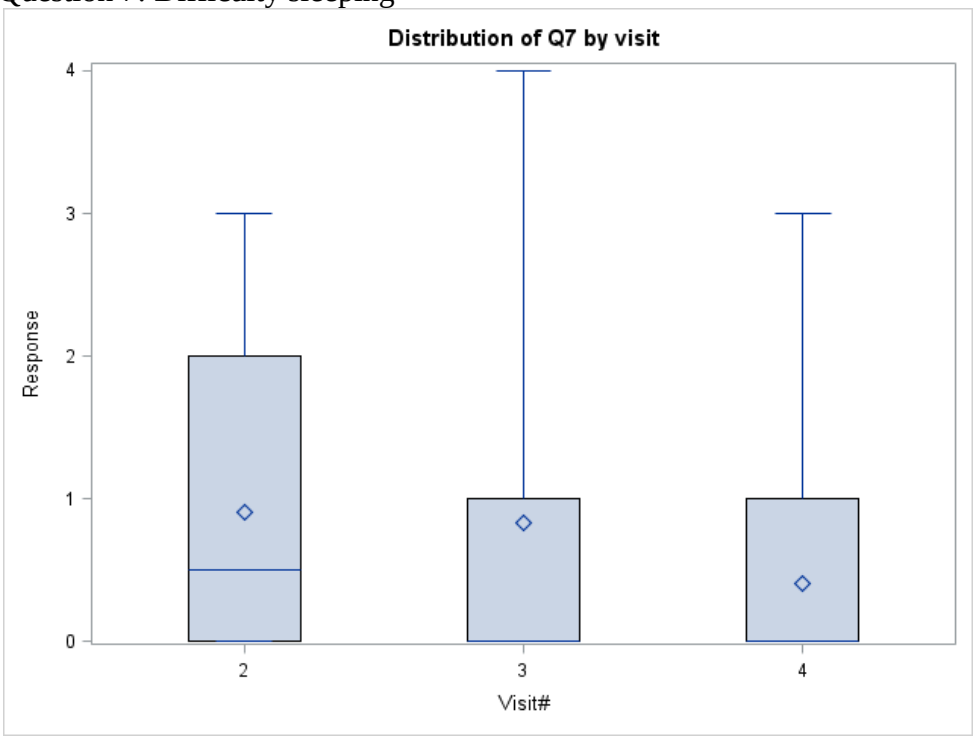


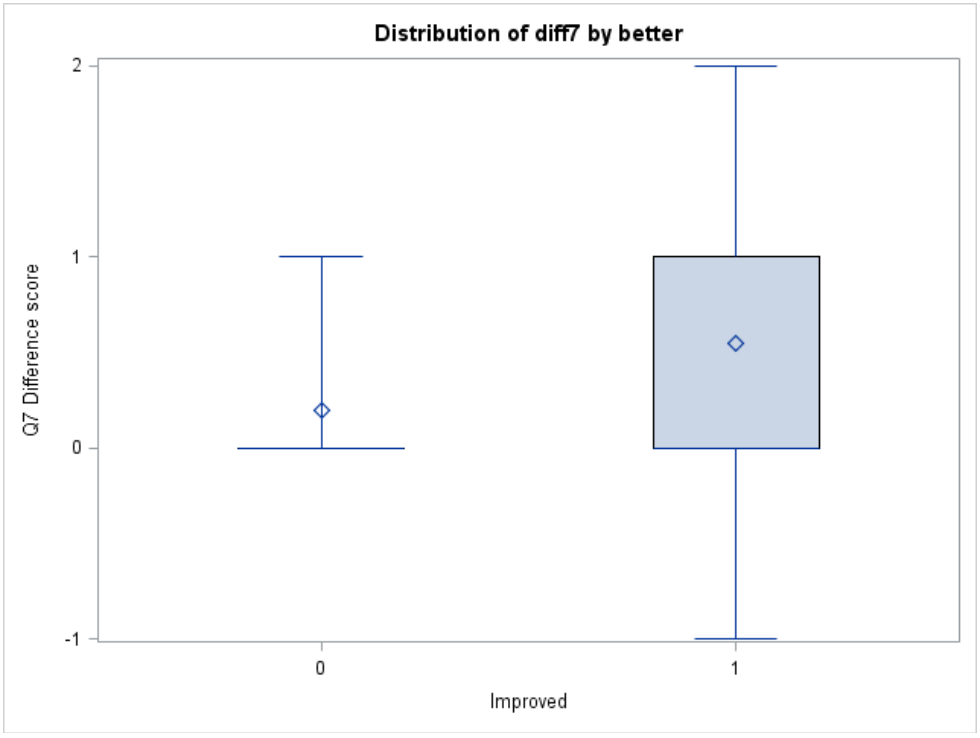
Question 6: Fatigue frequency



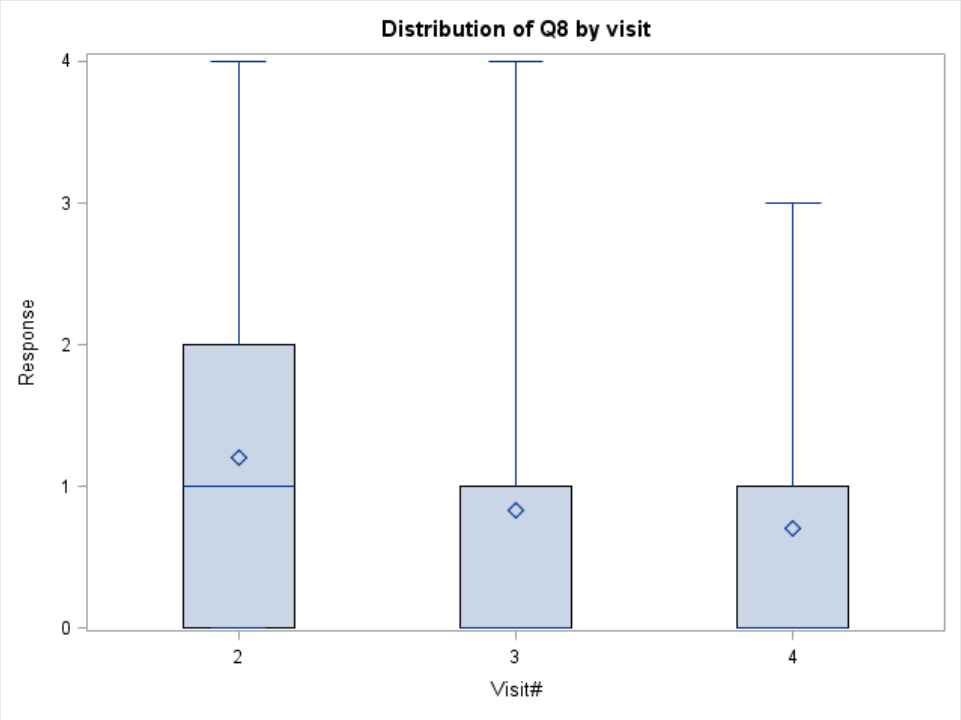


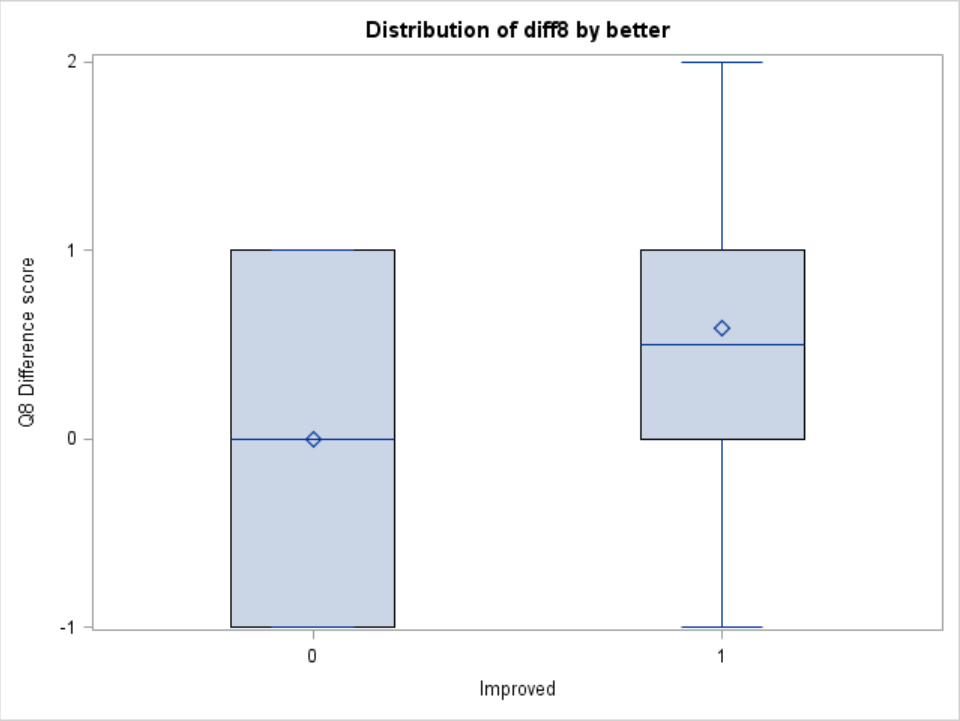
Question 7: Difficulty sleeping



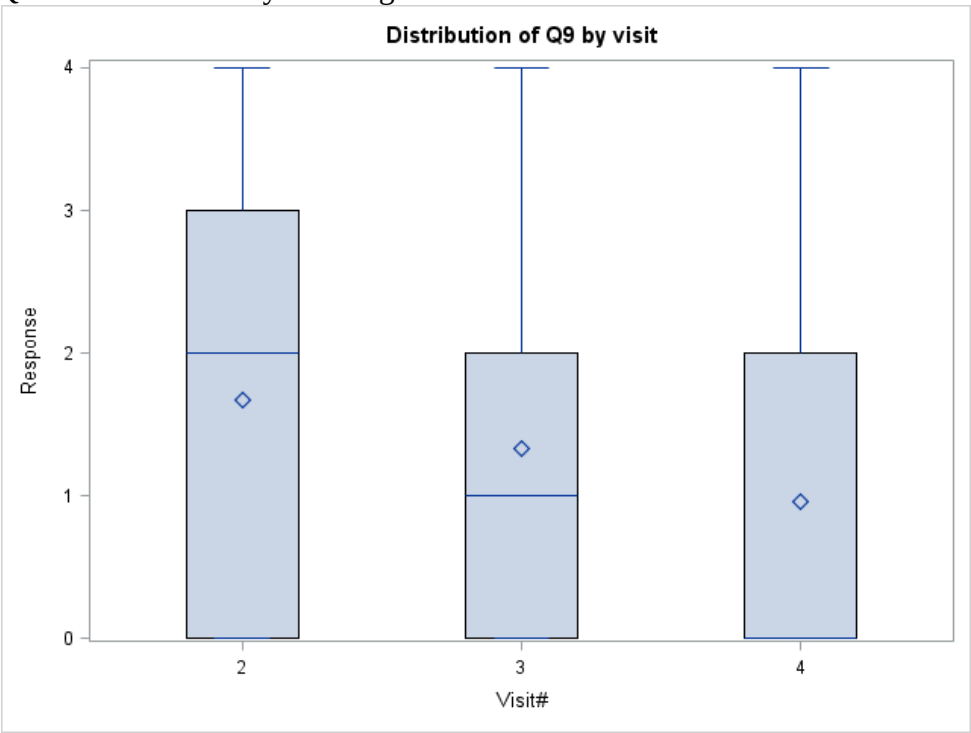


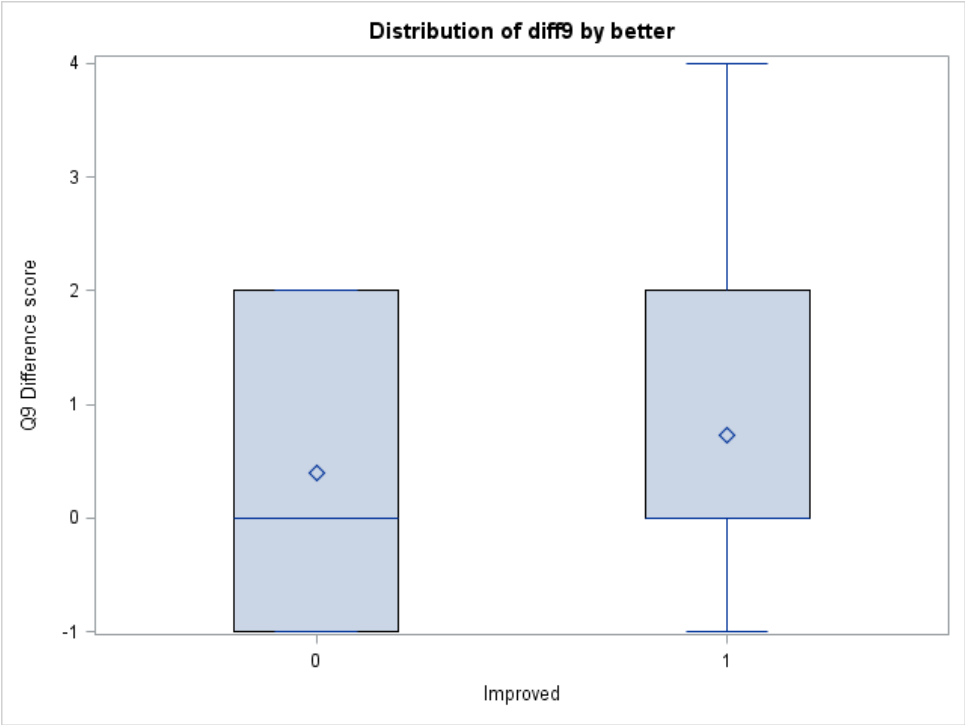
Question 8: Difficulty paying attention



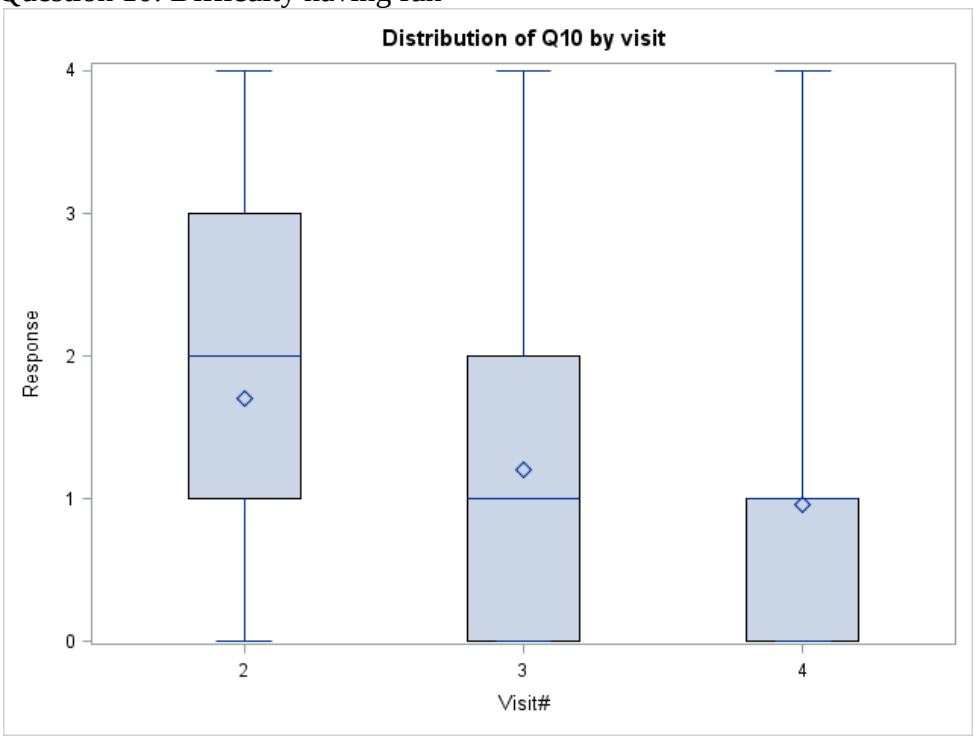


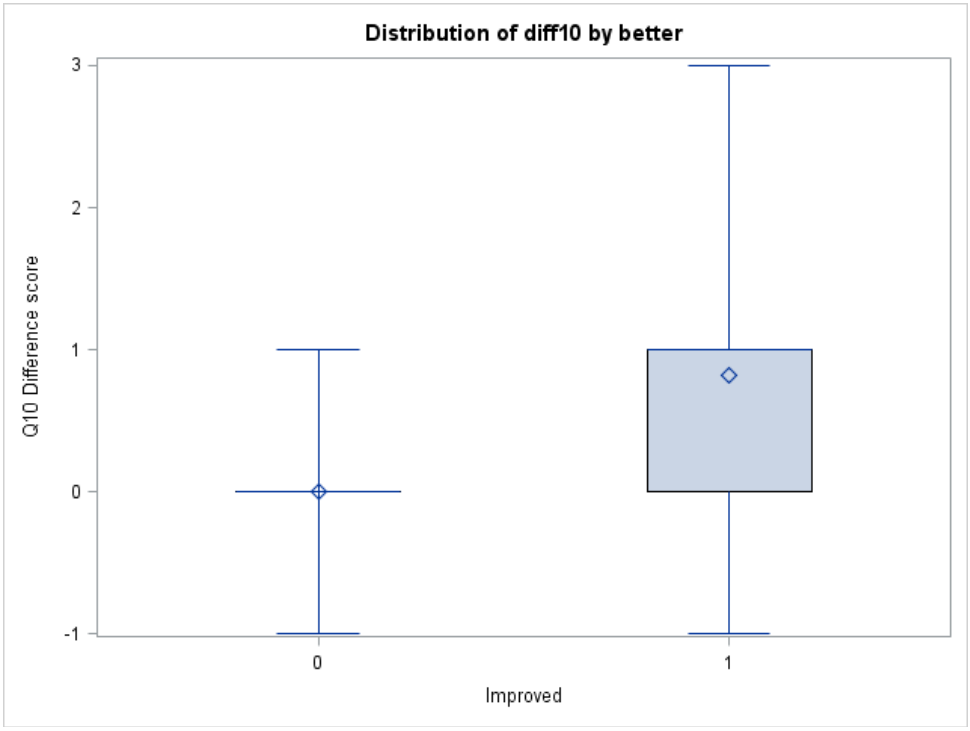
Question 9: Difficulty standing



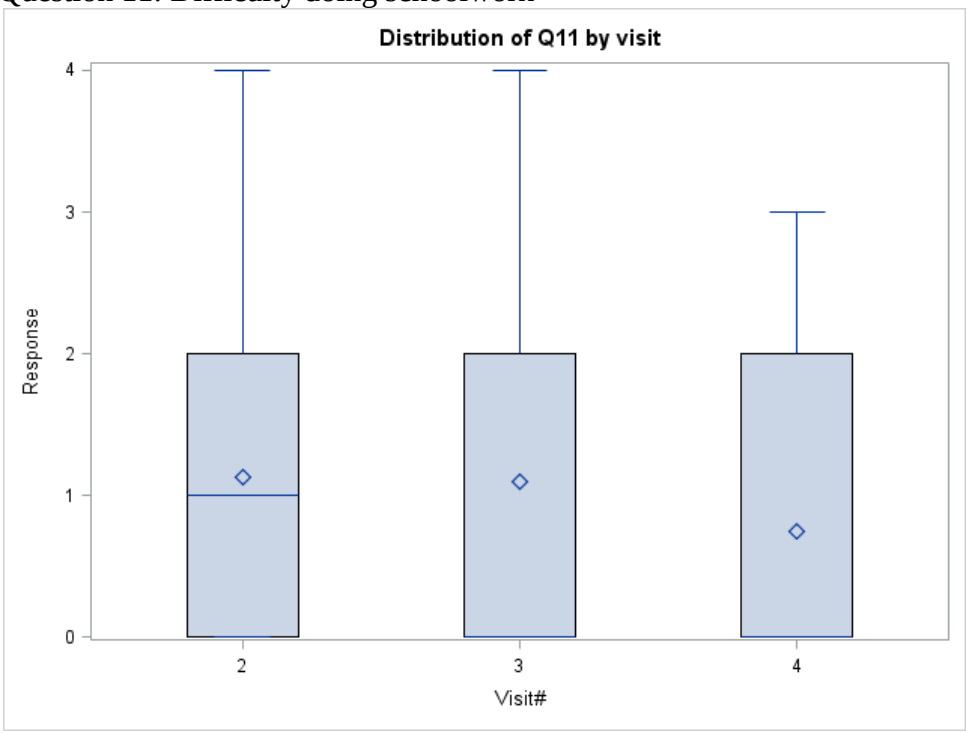


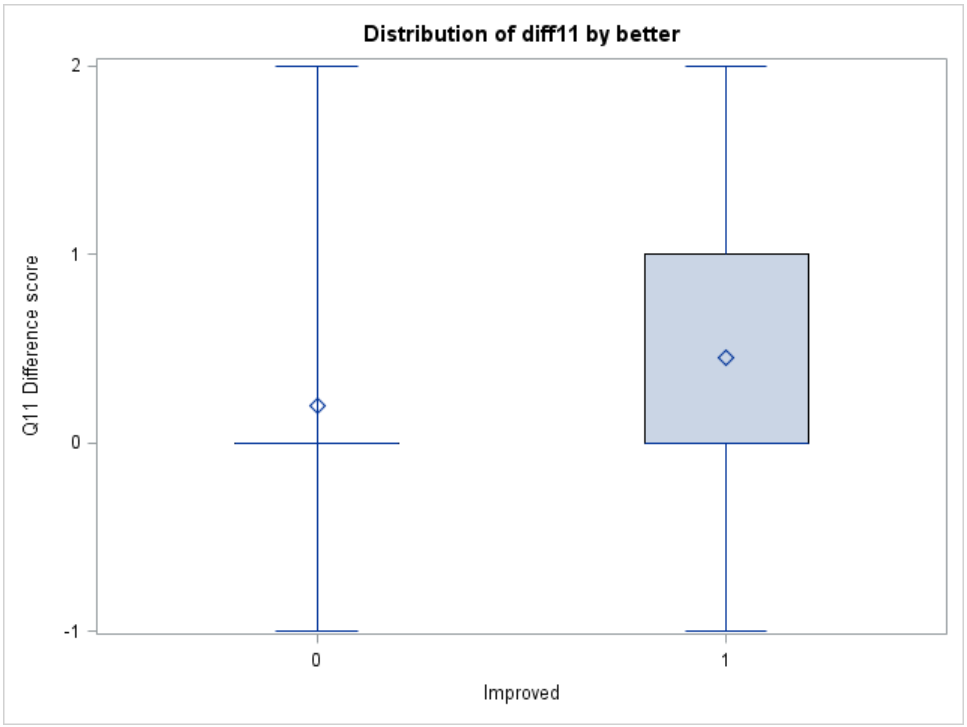
Question 10: Difficulty having fun



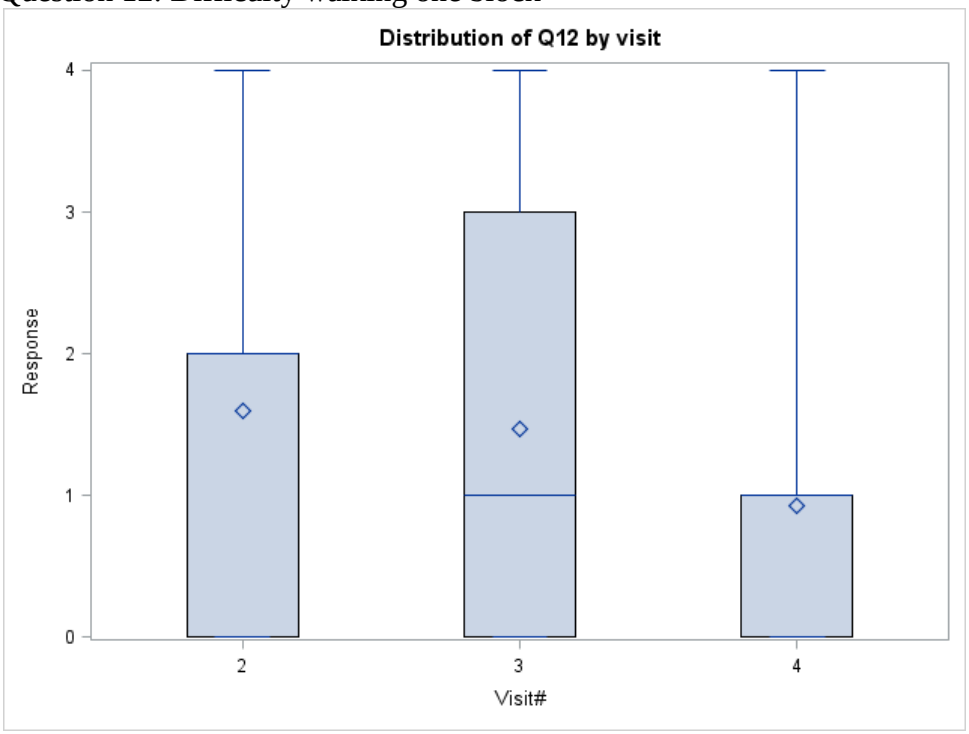


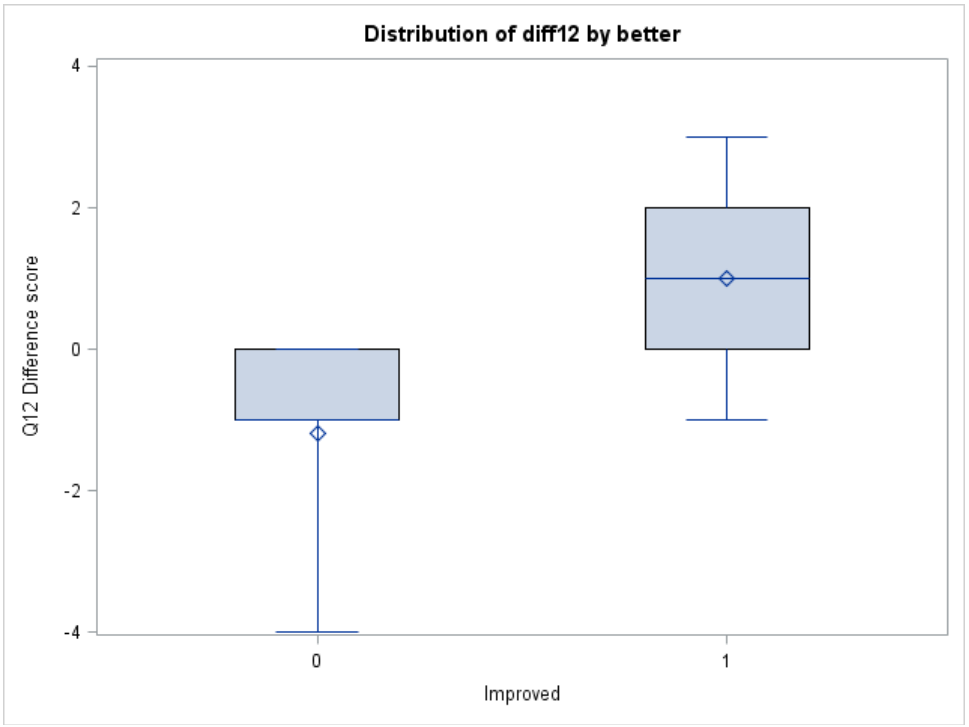
Question 11: Difficulty doing schoolwork



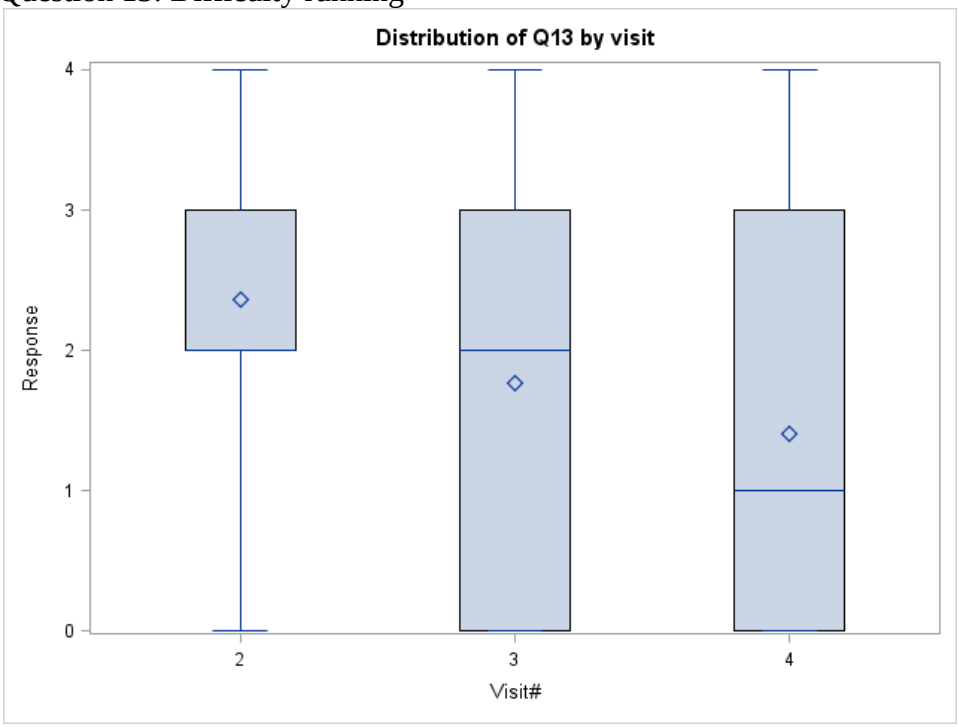


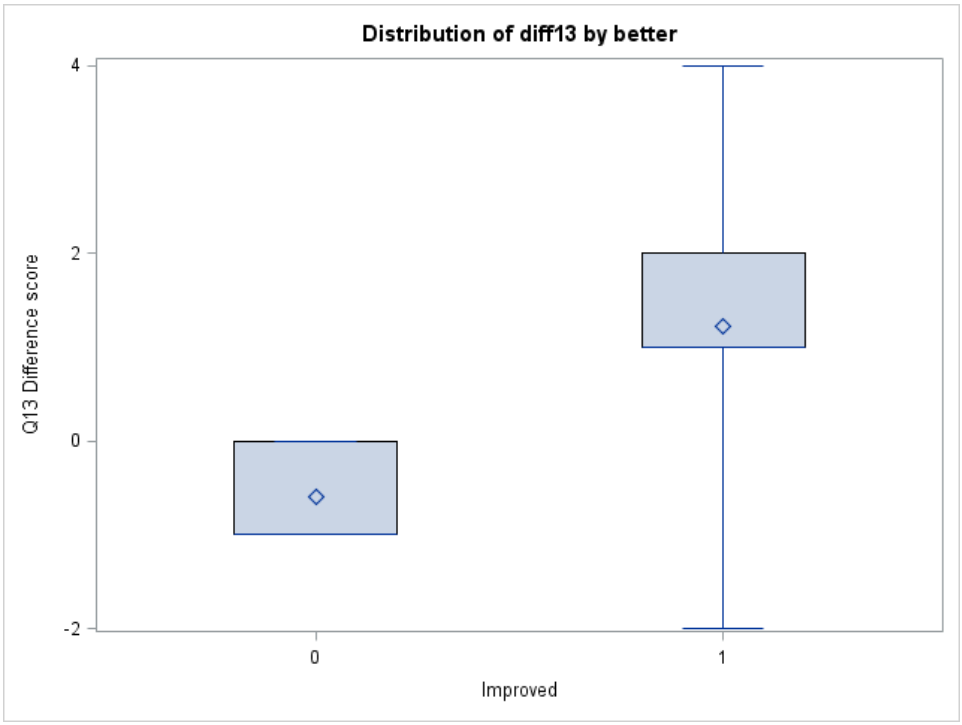
Question 12: Difficulty walking one block



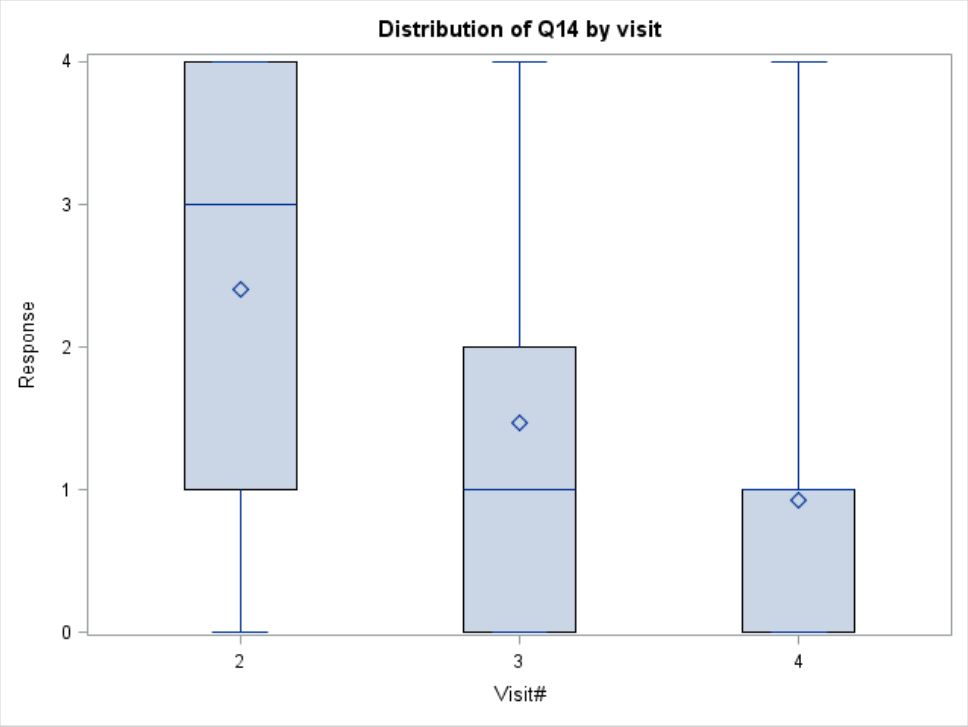


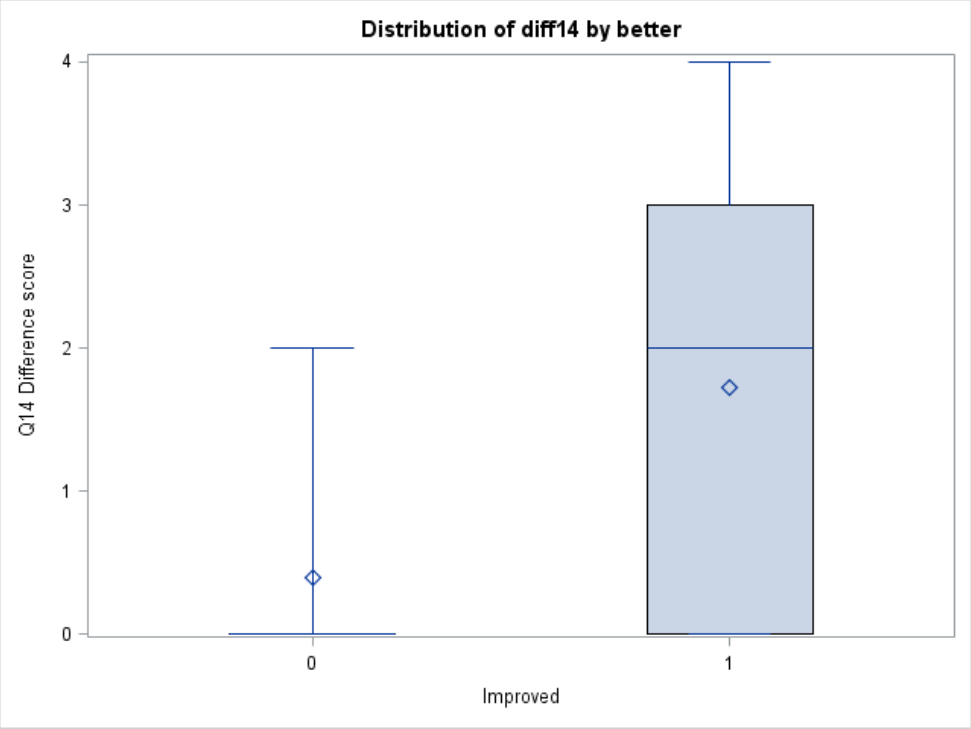
Question 13: Difficulty running



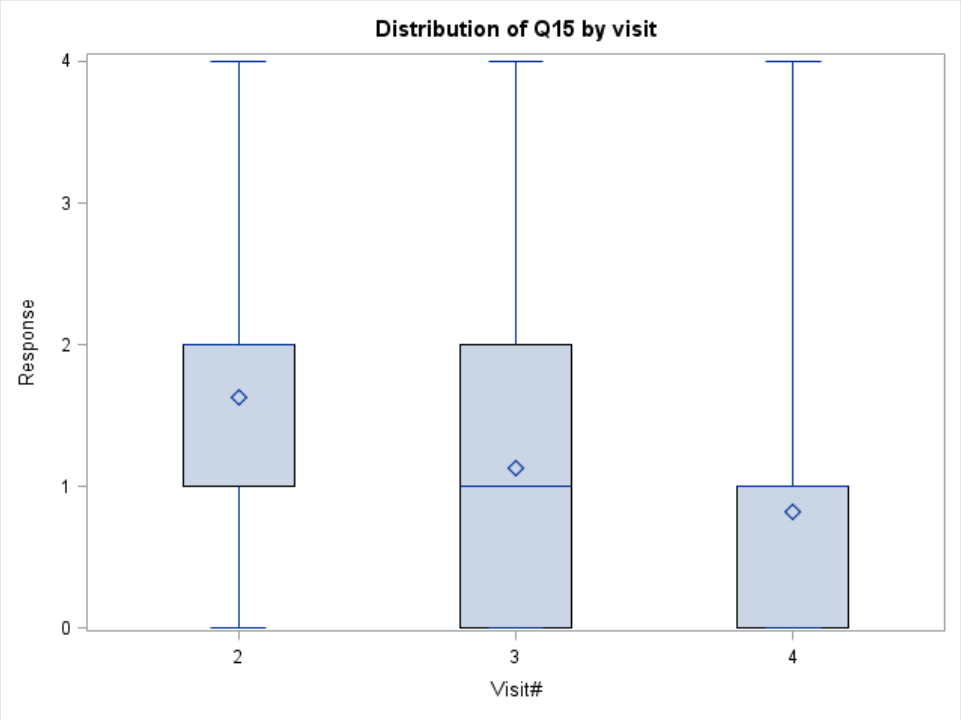


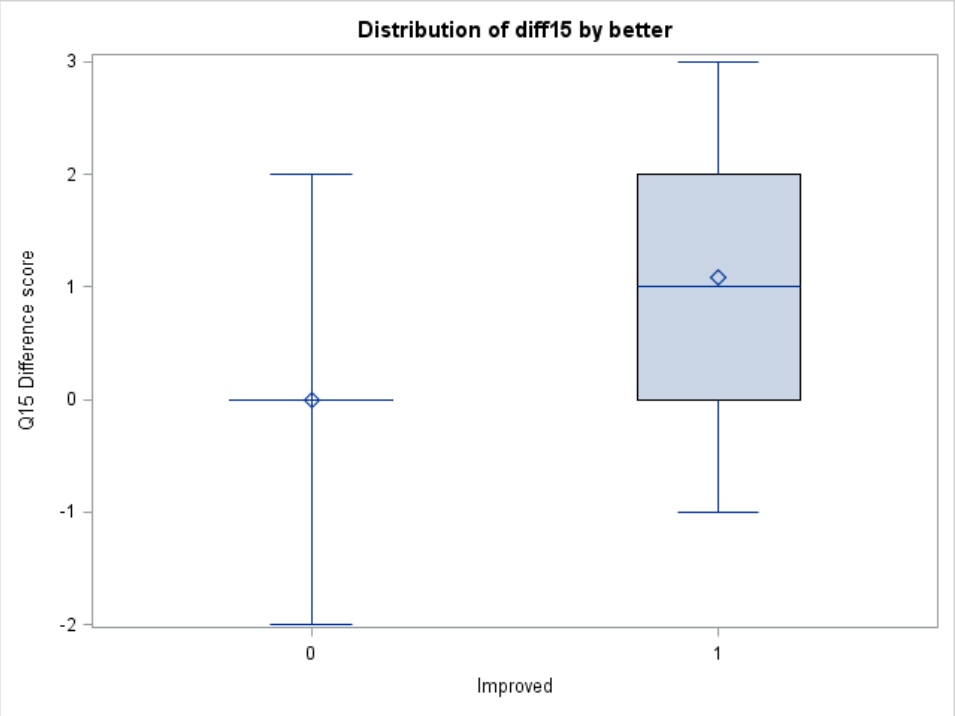
Question 14: Kept thinking how much wanted pain to stop



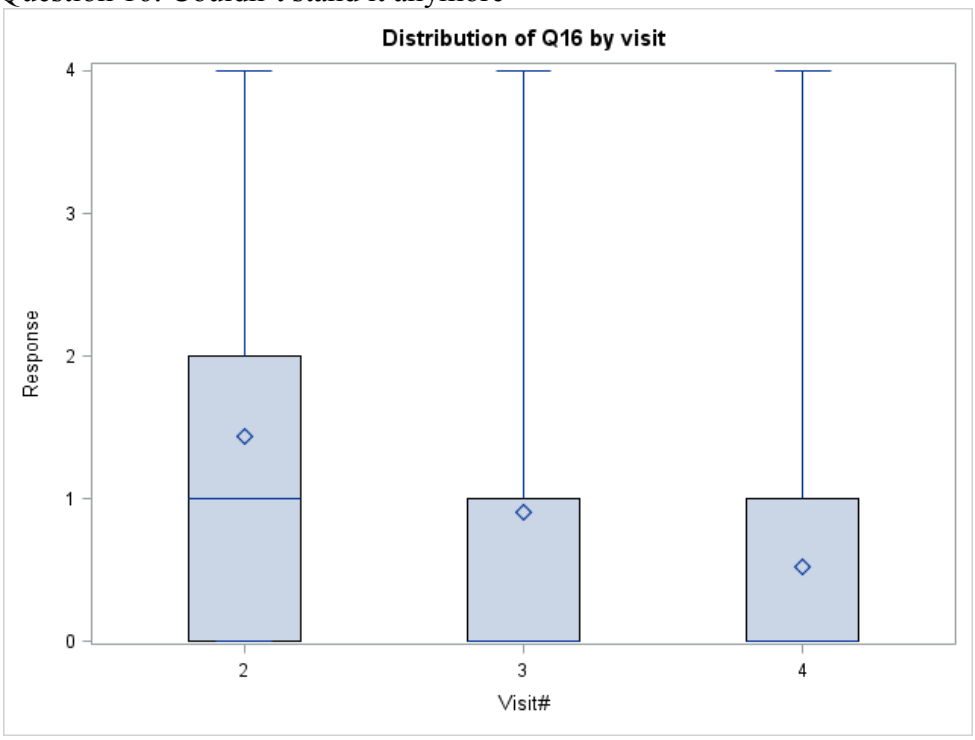


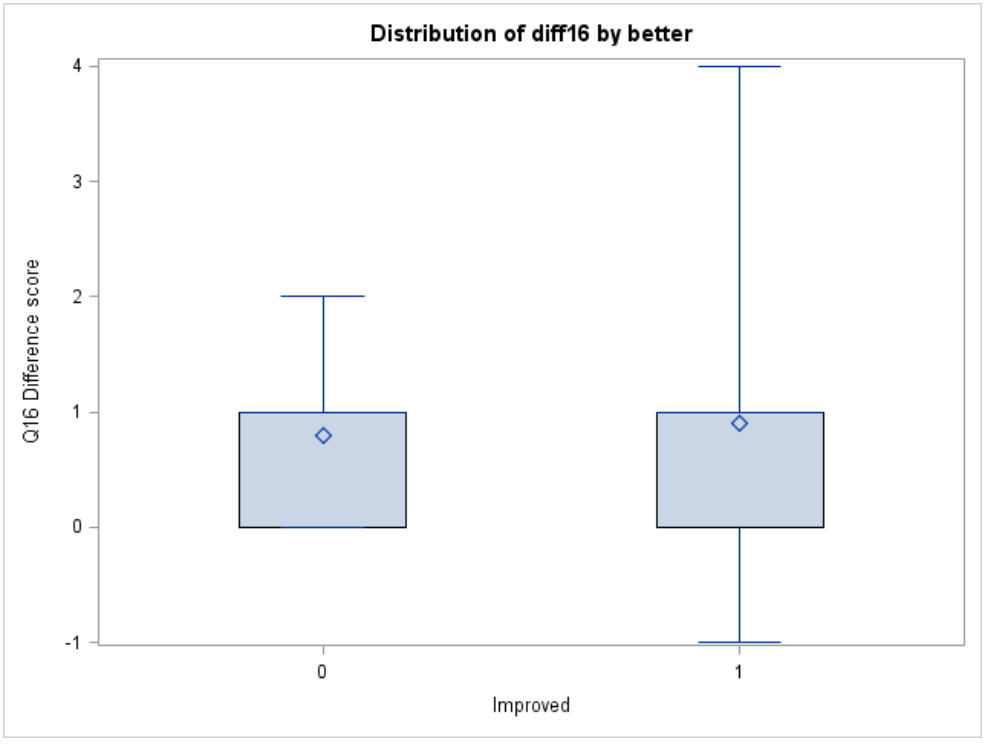
Question 15: Afraid pain would get worse



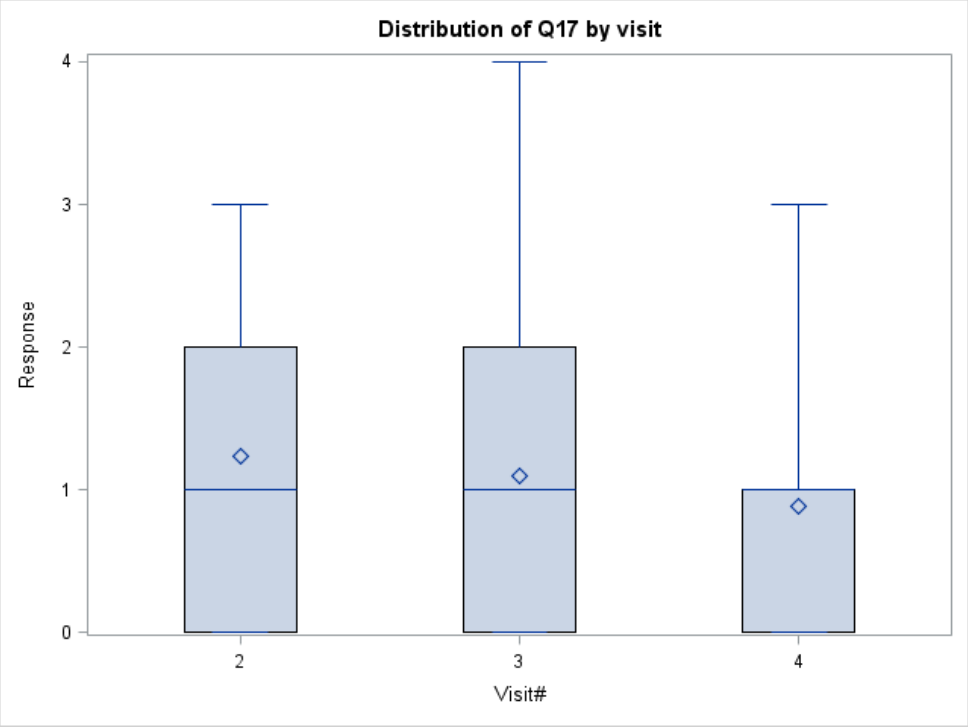


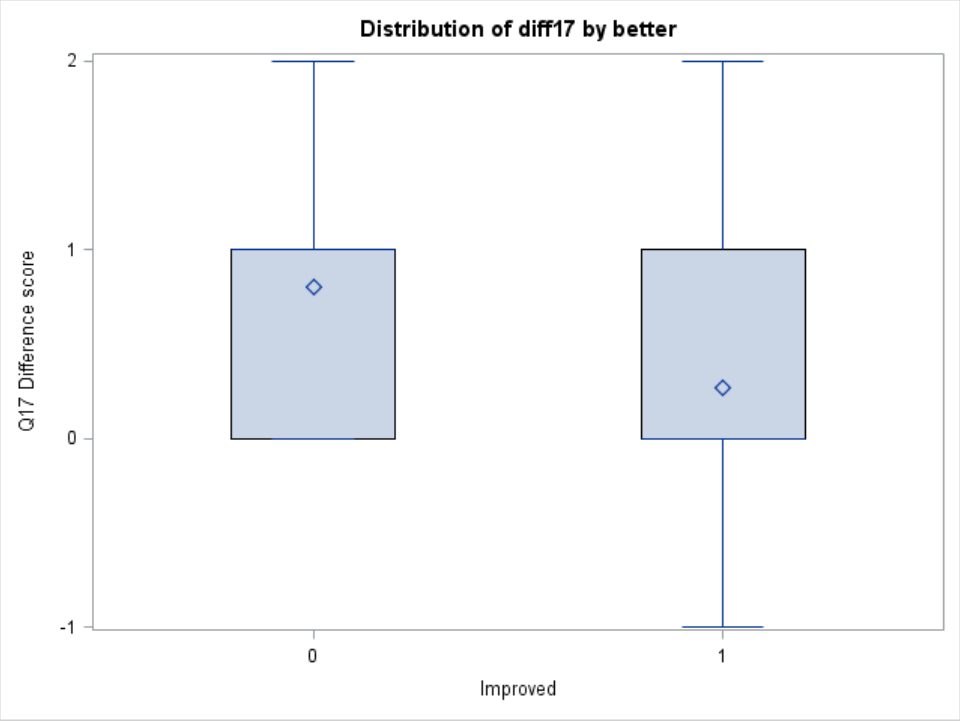
Question 16: Couldn't stand it anymore



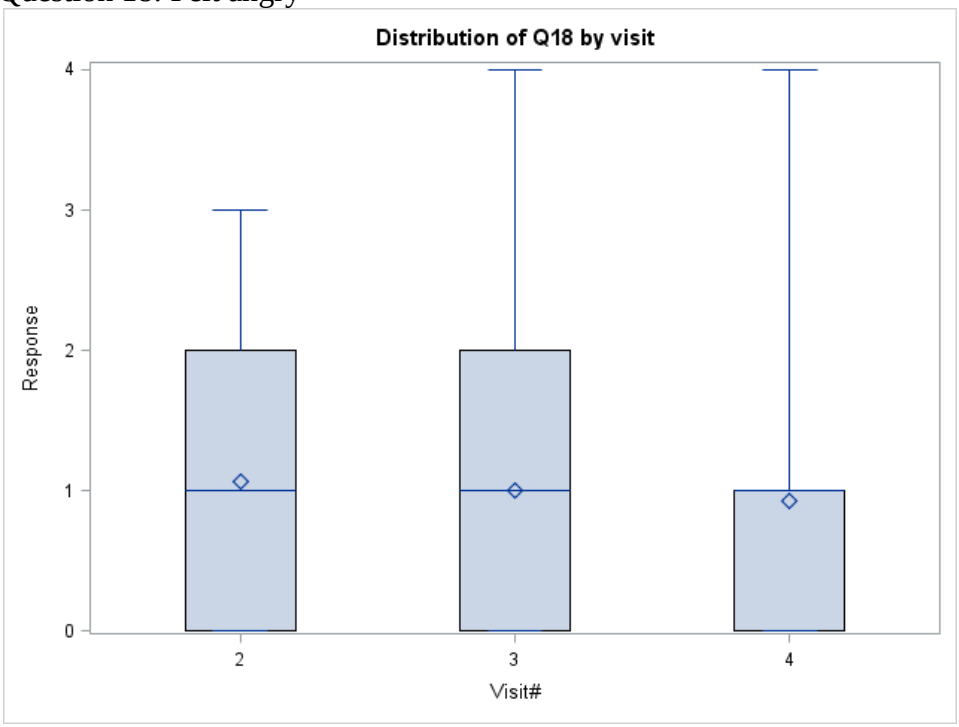


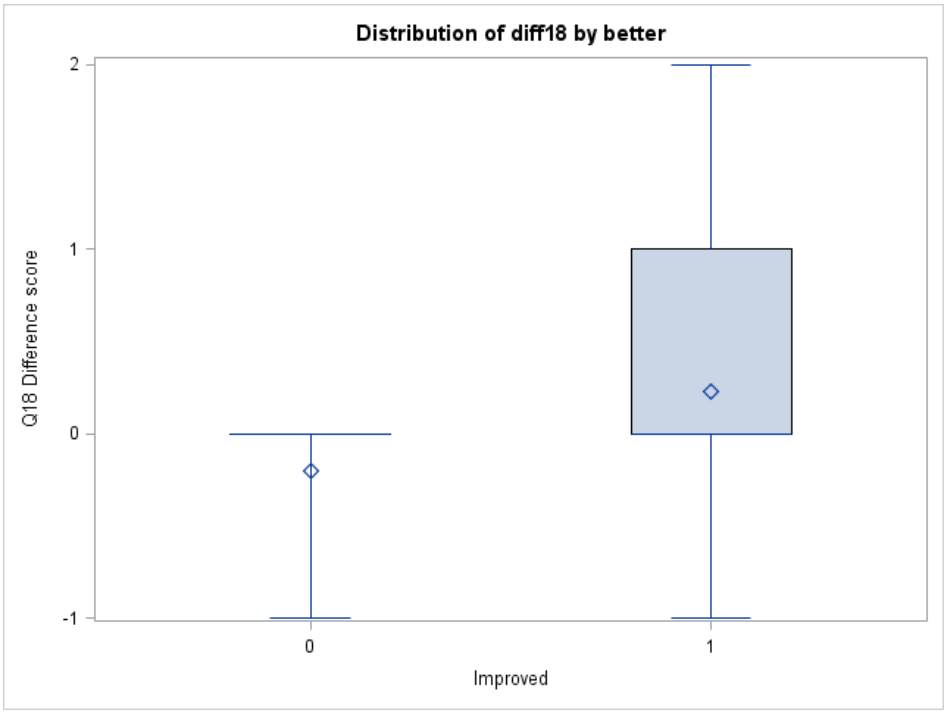
Question 17: Felt sad



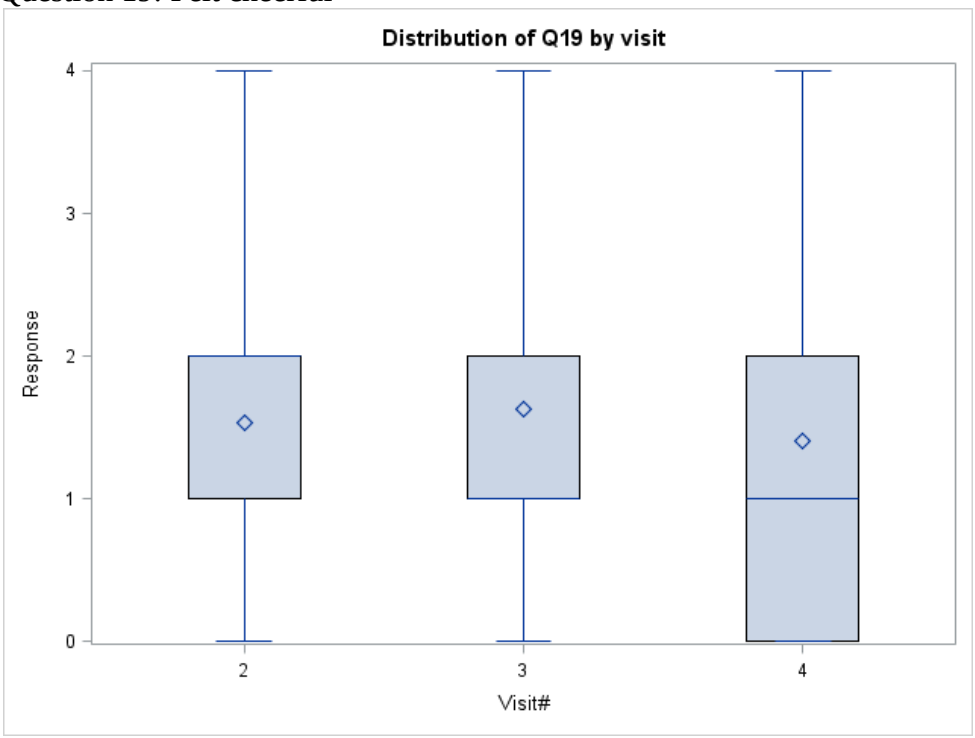


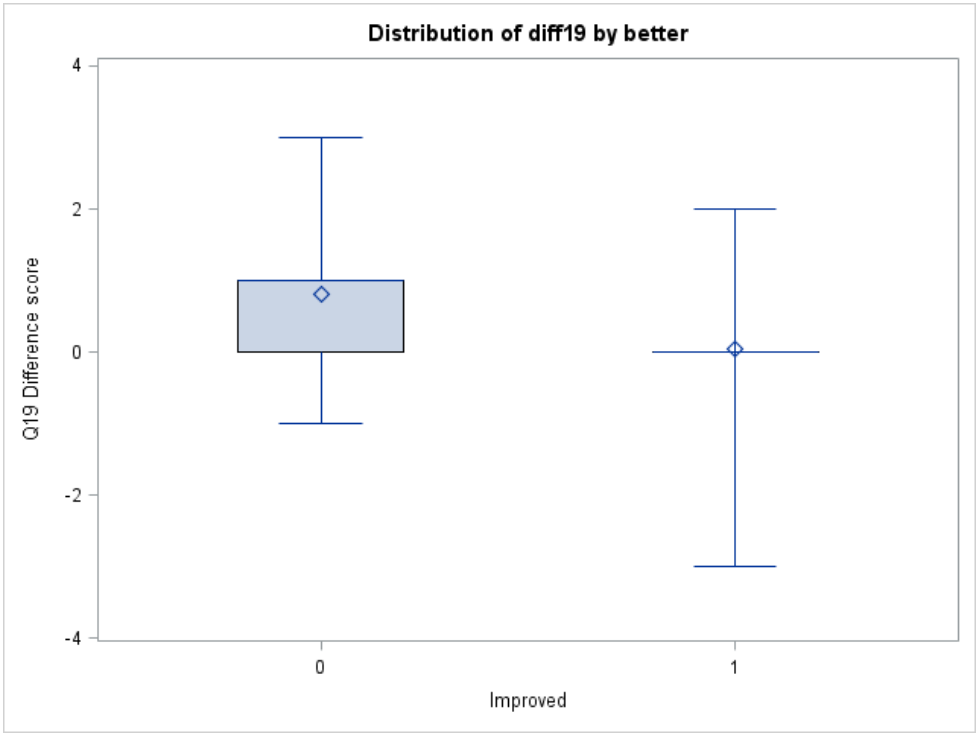
Question 18: Felt angry



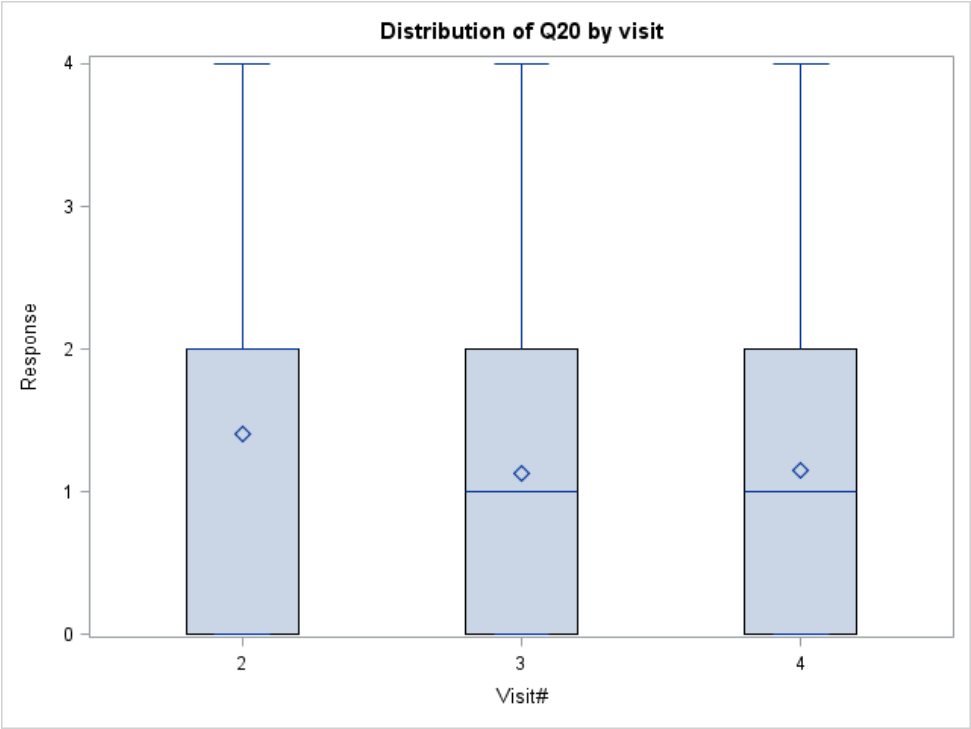


Question 19: Felt cheerful





Question 20: Felt worried



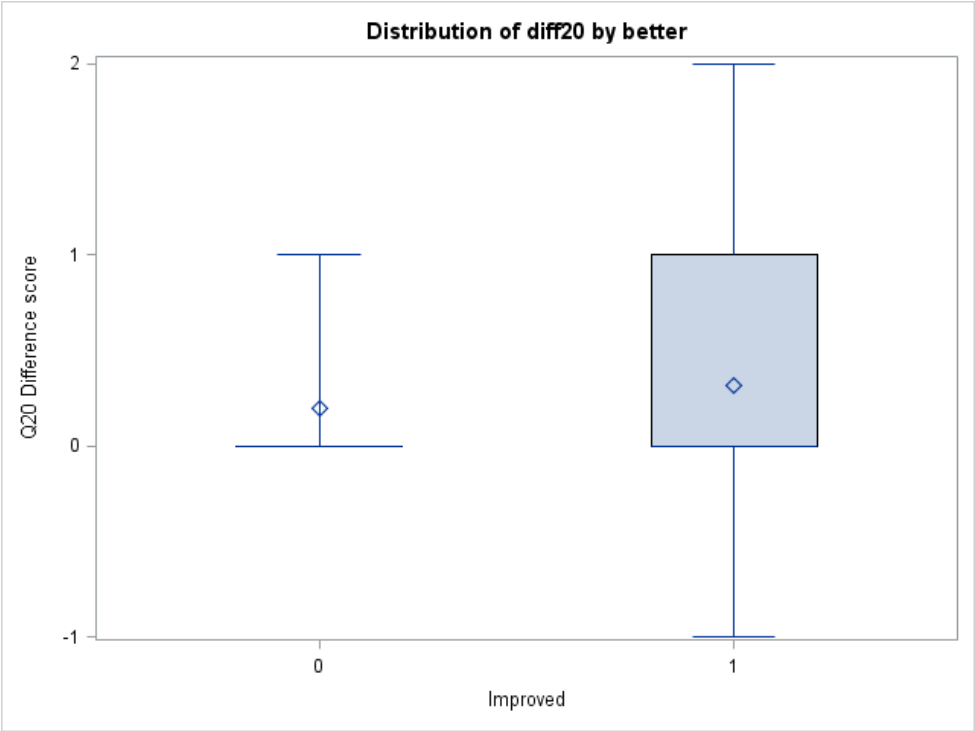
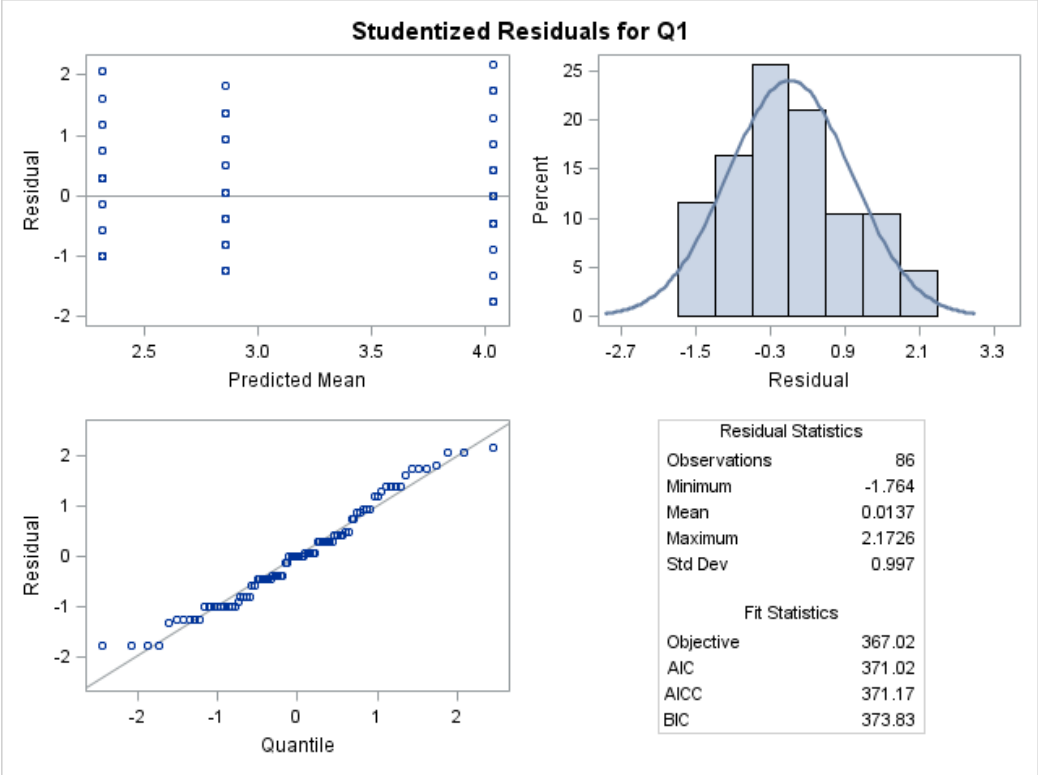


Table: The mean and median change for SUPER-KIDZ items in subjects reporting improvement (n=22) and those with no reported improvement (n=5) at 2 weeks post-injection

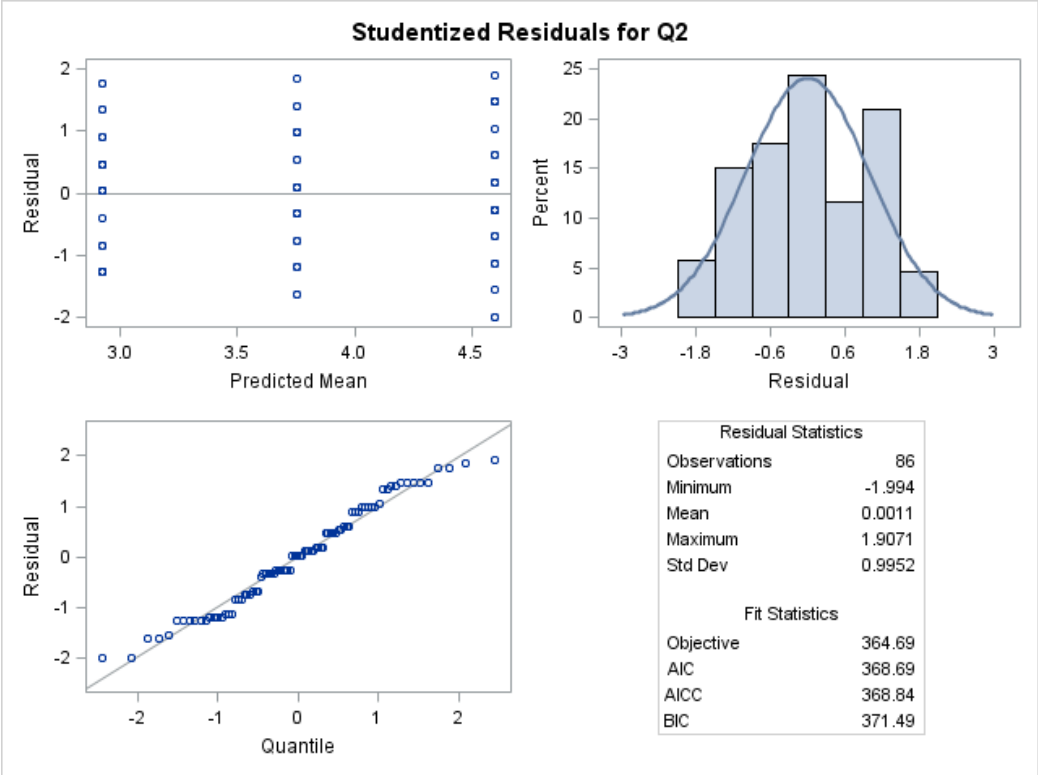
Item	Question	GRCP	Mean change score (SD)	Median change score (IQR)
Q1	Current pain	Improved	2.27 (2.16)	2.0 (1.0-4.0)
		Not improved	-0.20 (0.84)	0 (-1.0-0)
Q2	Average pain	Improved	2.05 (2.26)	2.0 (0-3.0)
		Not improved	0.20 (2.95)	0 (-1.0-0)
Q3	Pain frequency	Improved	1.32 (1.96)	1.0 (0-3.0)
		Not improved	-1.20 (1.30)	-1.0 (-2.0-0)
Q4	Pain duration	Improved	1.27 (1.58)	1.0 (0-2.0)
		Not improved	-0.20 (0.84)	0 (-1.0-0)
Q5	#Painful locations	Improved	2.45 (3.84)	2.0 (0-4.0)
		Not improved	1.60 (1.14)	2.0 (1.0-2.0)
Q6	Fatigue frequency	Improved	0.14 (0.94)	0 (0-1.0)
		Not improved	0.40 (0.55)	0 (0-1.0)
Q7	Sleeping	Improved	0.54 (0.86)	0 (0-1.0)
		Not improved	0.20 (0.45)	0 (0-0)
Q8	Paying attention	Improved	0.59 (0.80)	0.5 (0-1.0)
		Not improved	0 (1.00)	0 (-1.0-1.0)
Q9	Standing	Improved	0.73 (1.16)	0 (0-2.0)
		Not improved	0.40 (1.52)	0 (-1.0-2.0)
Q10	Having fun	Improved	0.82 (0.96)	1.0 (0-1.0)
		Not improved	0 (0.71)	0 (0-0)
Q11	Schoolwork	Improved	0.45 (0.91)	0 (0-1.0)
		Not improved	0.20 (1.10)	0 (0-0)
Q12	Walk 1 block	Improved	1.00 (1.02)	1.0 (0-2.0)
		Not improved	-1.20 (1.64)	-1.0 (-1.0-0)
Q13	Running	Improved	1.23 (1.23)	1.0 (1.0-2.0)
		Not improved	-0.60 (0.55)	-1.0 (-1.0-0)
Q14	Want pain to stop	Improved	1.73 (1.35)	2.0 (0-3.0)
		Not improved	0.40 (0.89)	0 (0-0)
Q15	Afraid pain worse	Improved	1.09 (1.02)	1.0 (0-2.0)
		Not improved	0 (1.41)	0 (0-0)
Q16	Can’t stand it	Improved	0.91 (1.15)	1.0 (0-1.0)
		Not improved	0.80 (0.84)	1.0 (0-1.0)
Q17	Feel sad	Improved	0.27 (0.70)	0 (0-1.0)
		Not improved	0.80 (0.84)	1.0 (0-1.0)
Q18	Feel angry	Improved	0.23 (0.69)	0 (0-1.0)
		Not improved	-0.20 (0.45)	0 (0-0)
Q19	Feel cheerful	Improved	0.05 (1.17)	0 (0-0)
		Not improved	0.80 (1.48)	1.0 (0-1.0)
Q20	Feel worried	Improved	0.32 (0.78)	0 (0-1.0)
		Not improved	0.20 (0.45)	0 (0-0)

9. Residual plots for linear mixed models: Homoscedasticity and normality of the residuals

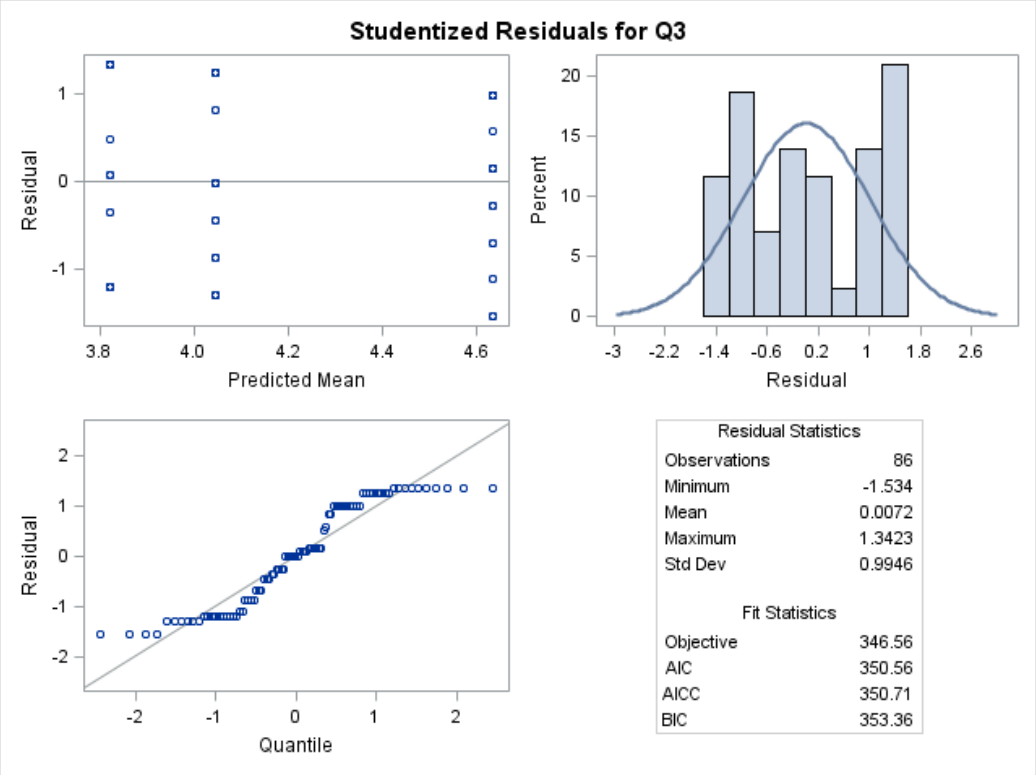
Question 1: Current pain intensity



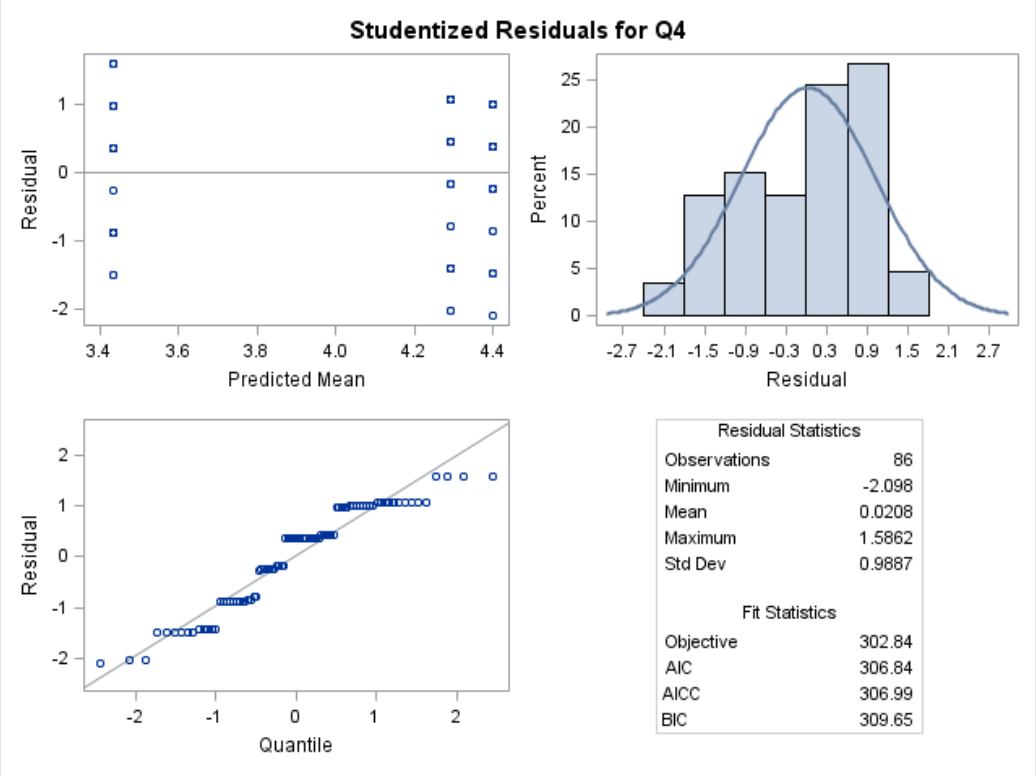
Question 2: Average pain intensity over past week



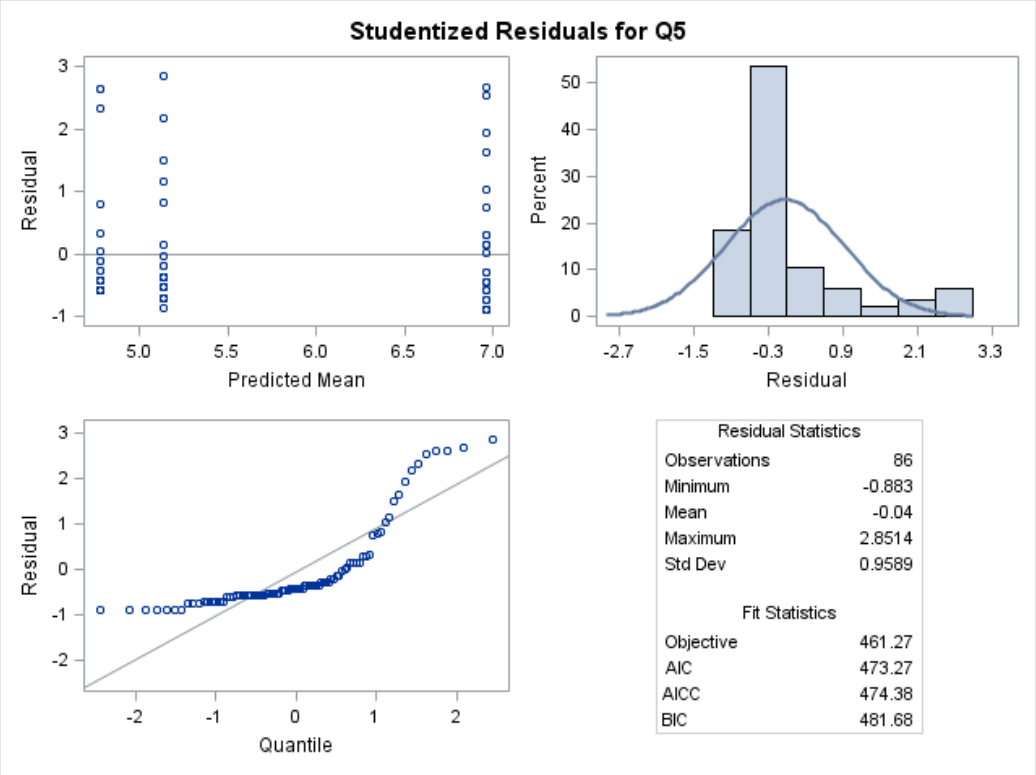
Question 3: Frequency of pain



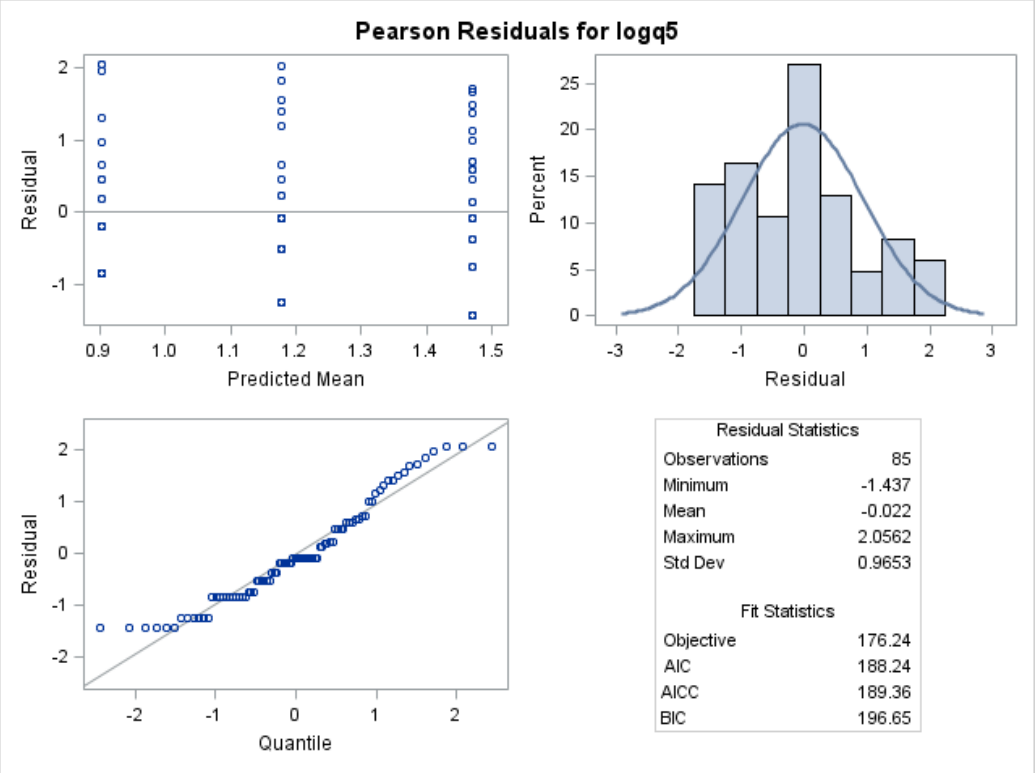
Question 4: Duration of pain



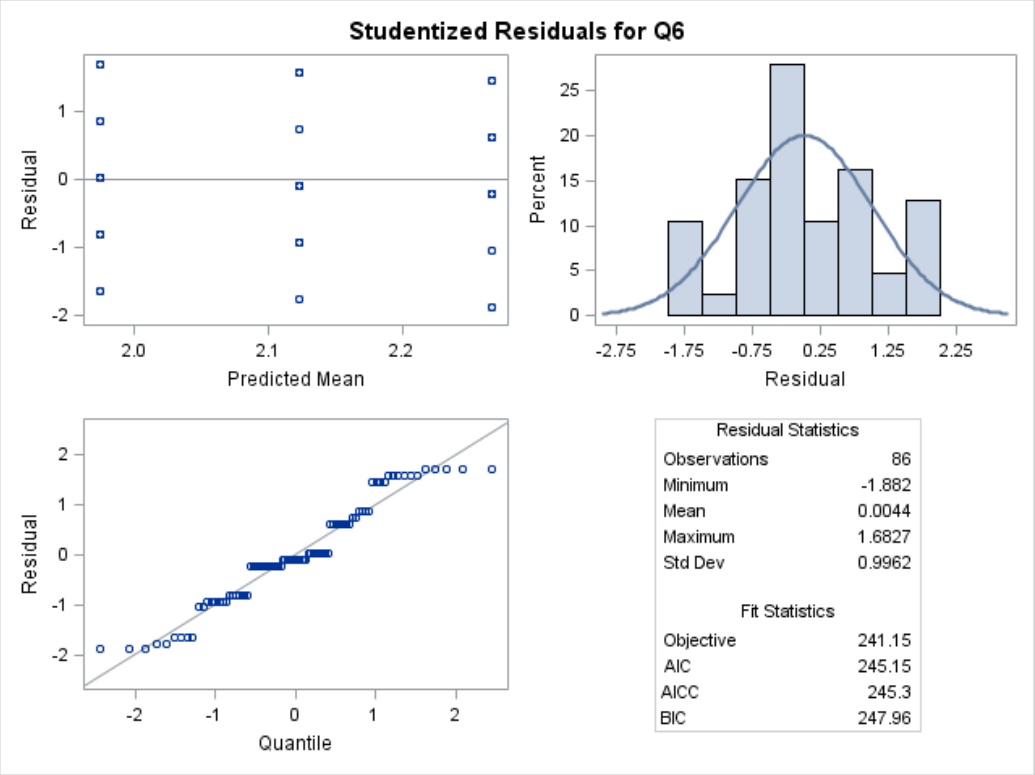
Question 5: Number of painful body locations (original model)



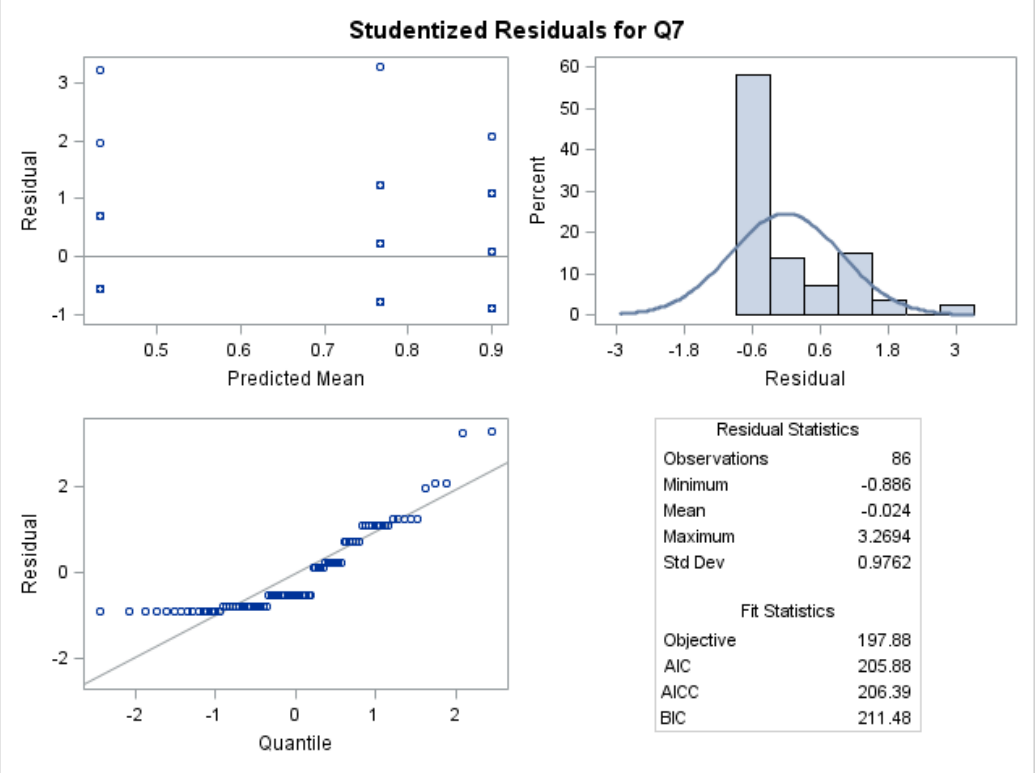
Question 5: Number of painful body locations (log-transformed model). # Obs=85 because 1 patient reported 0 painful locations.



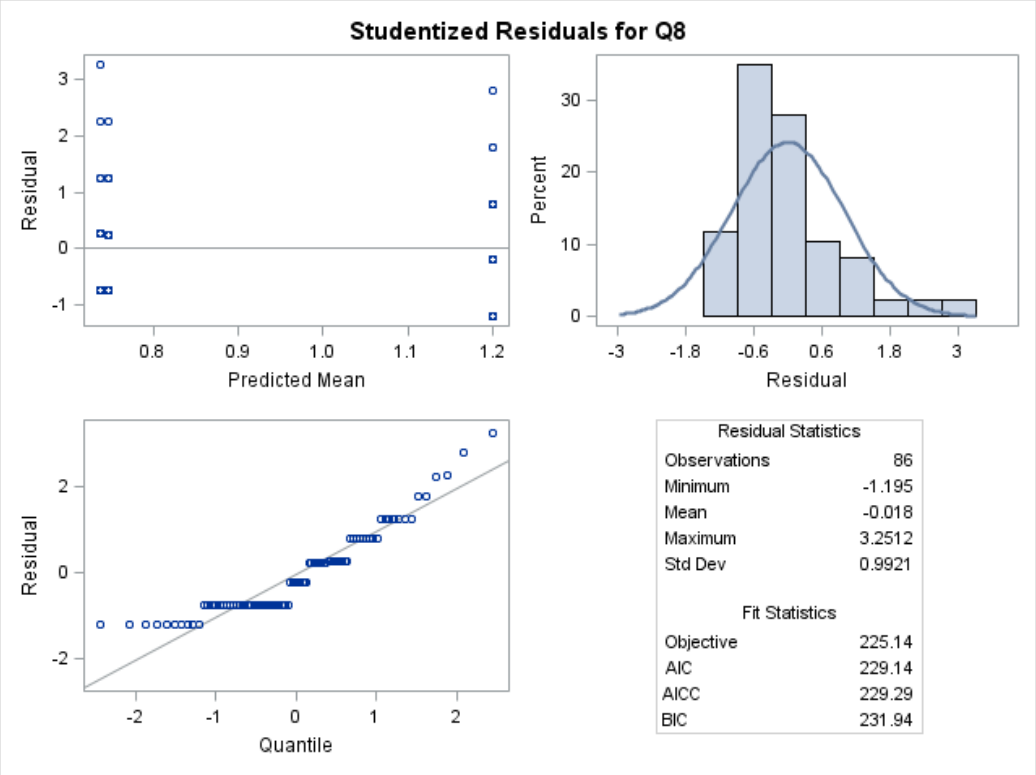
Question 6: Fatigue frequency



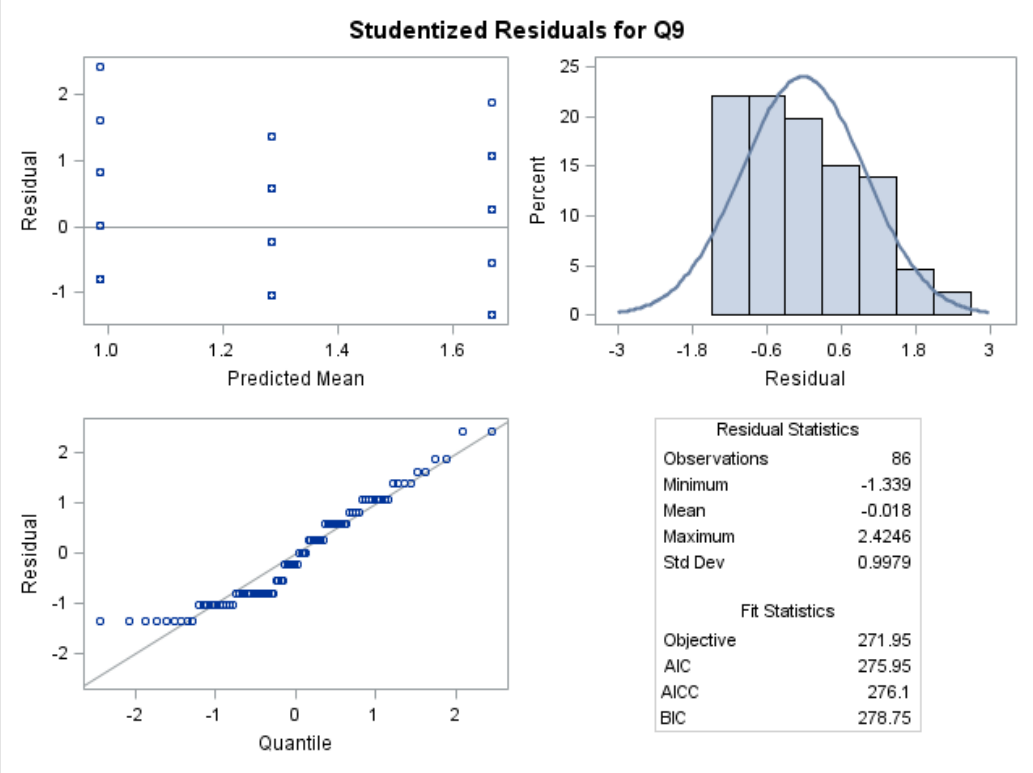
Question 7: Difficulty sleeping



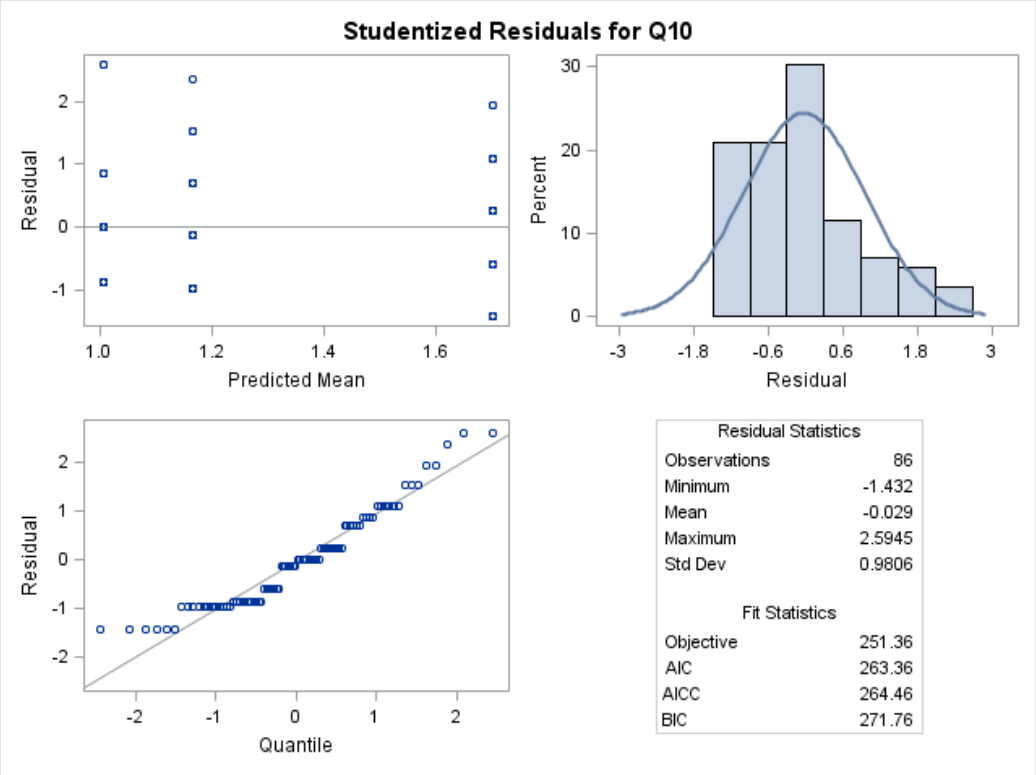
Question 8: Difficulty paying attention



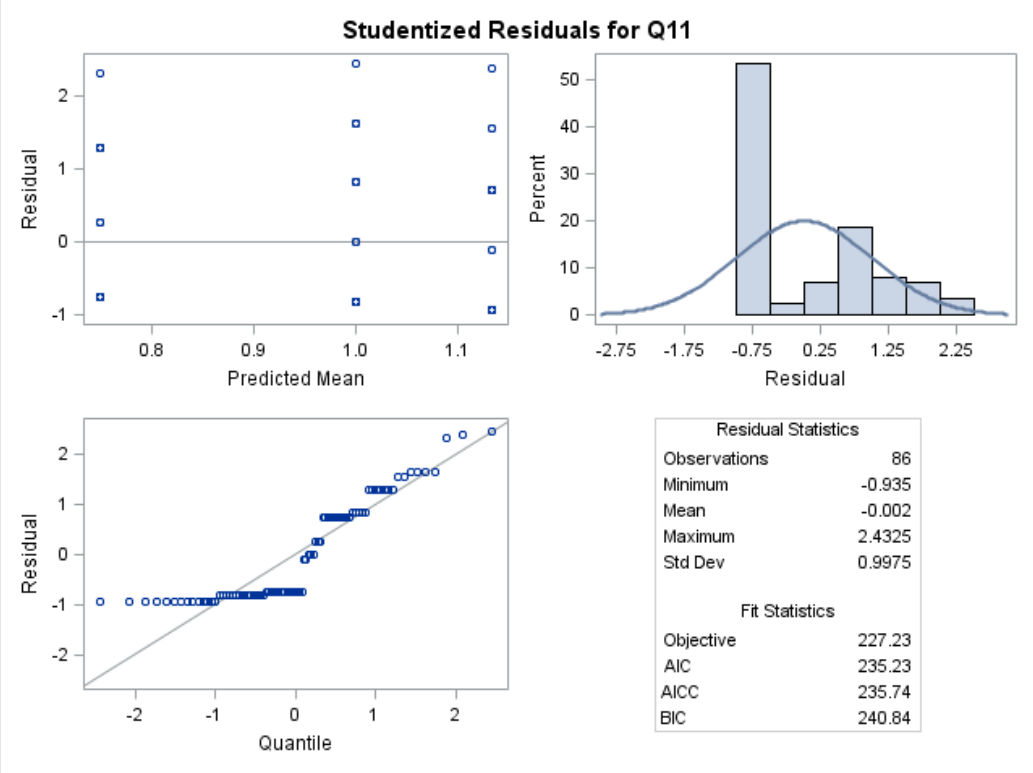
Question 9: Difficulty standing



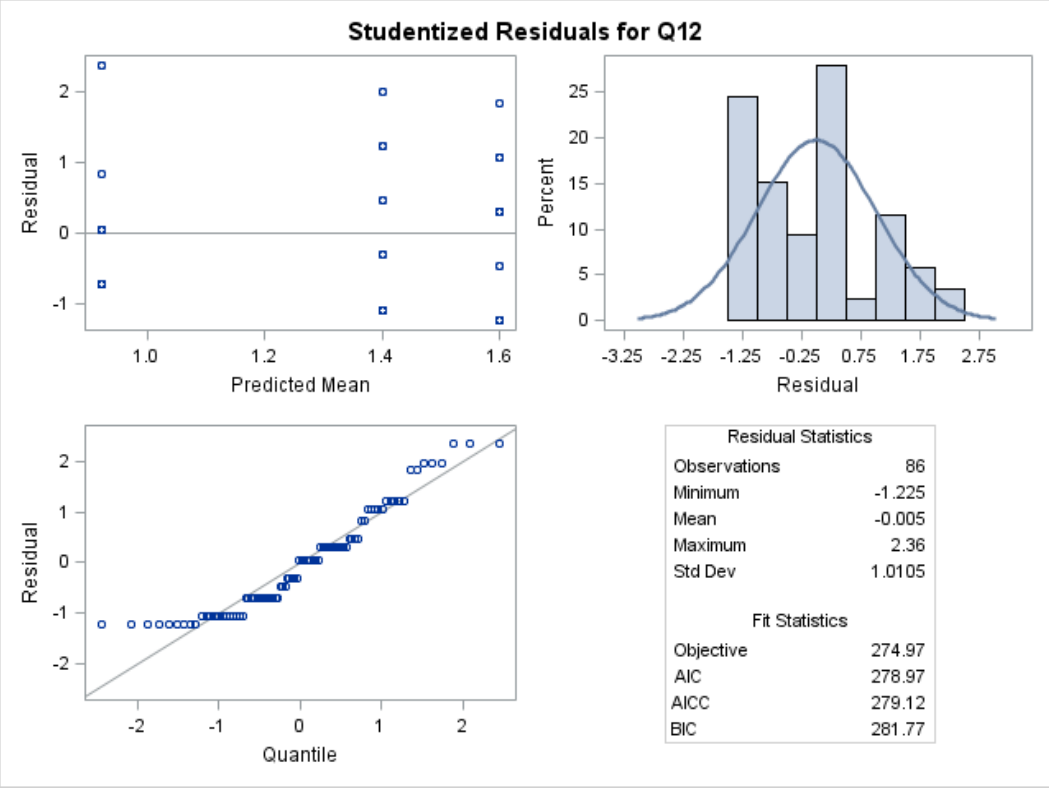
Question 10: Difficulty having fun



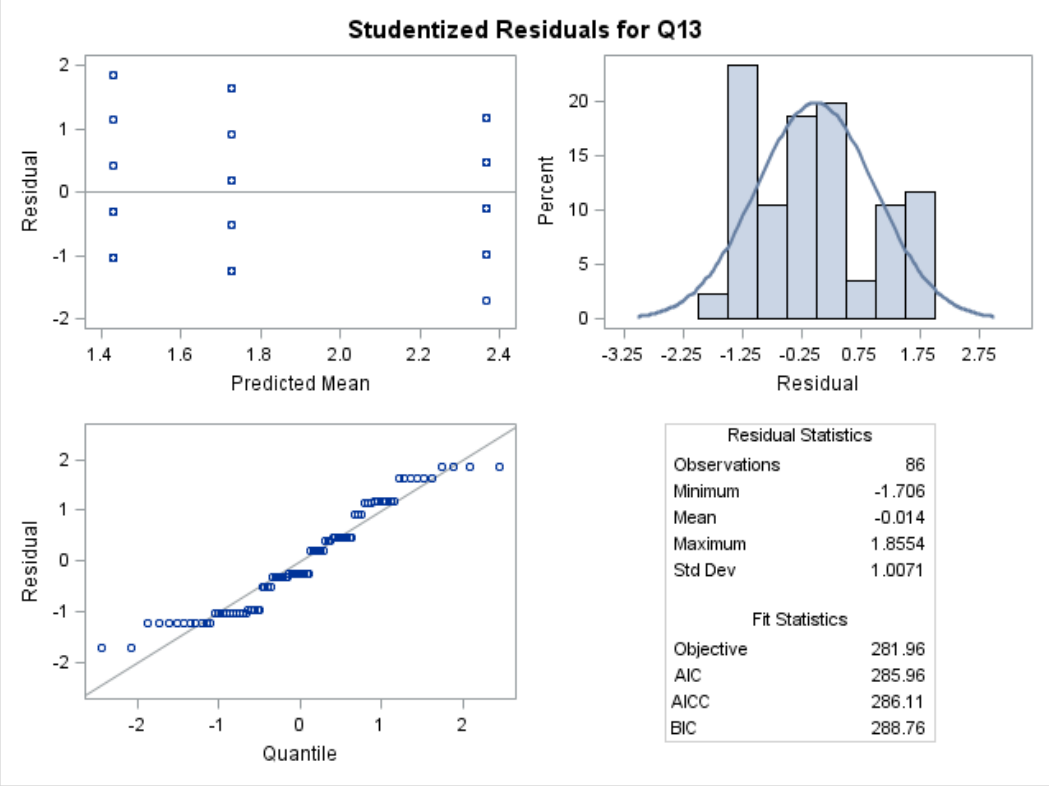
Question 11: Difficulty doing schoolwork



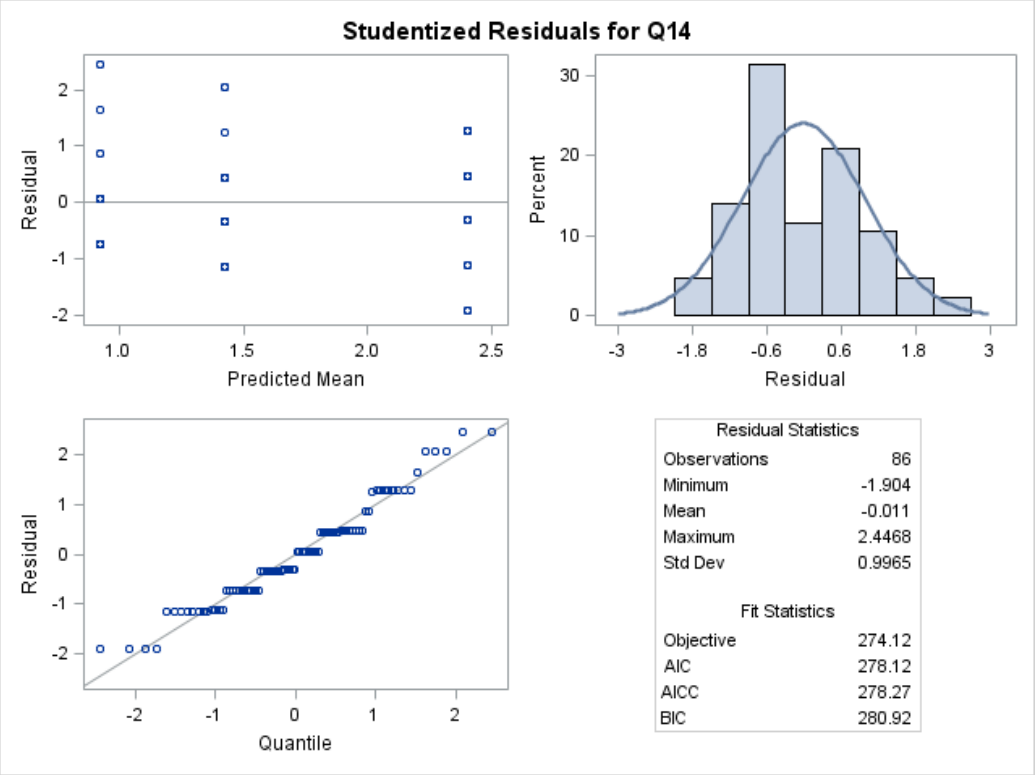
Question 12: Difficulty walking one block



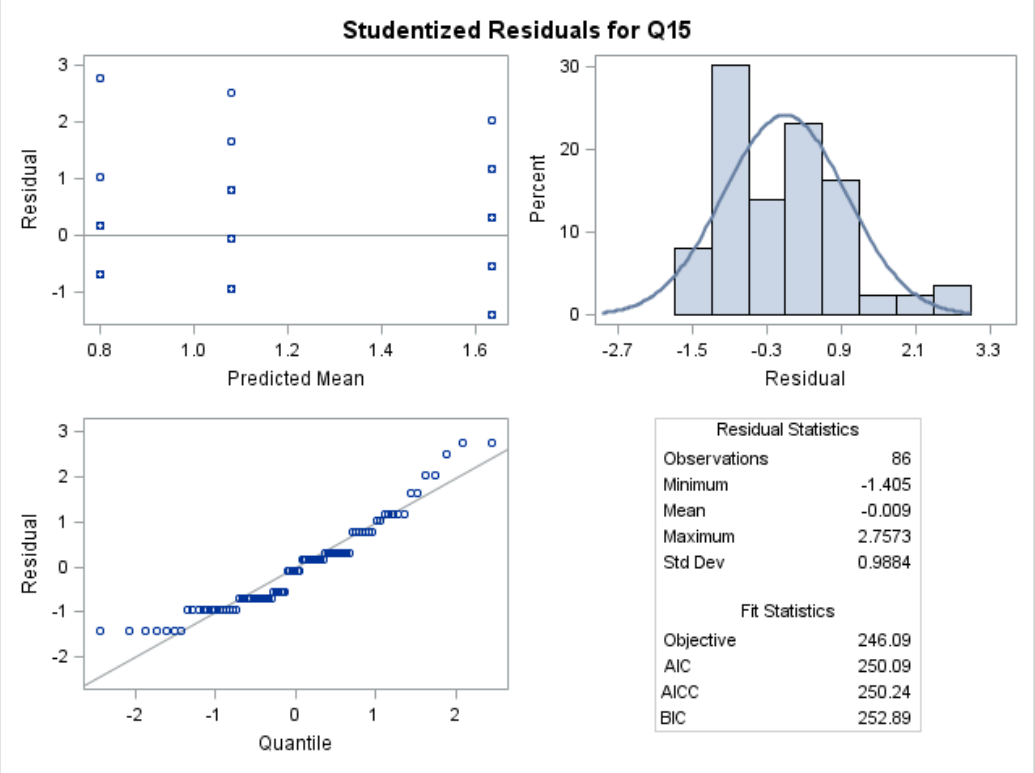
Question 13: Difficulty running



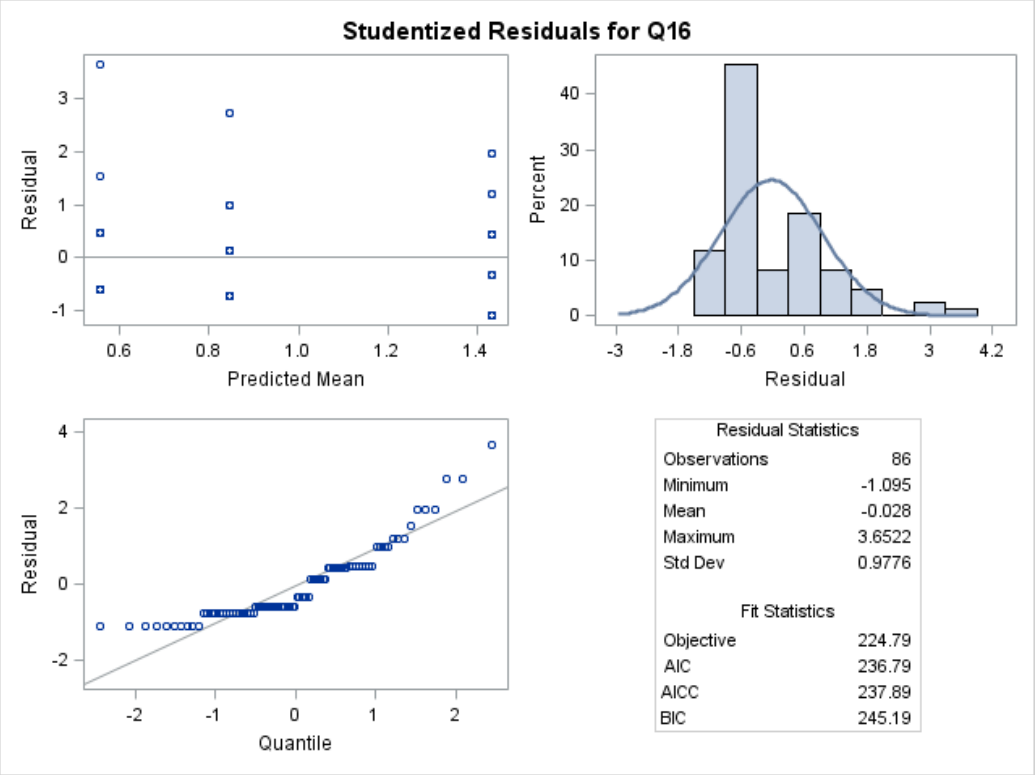
Question 14: Kept thinking how much wanted pain to stop



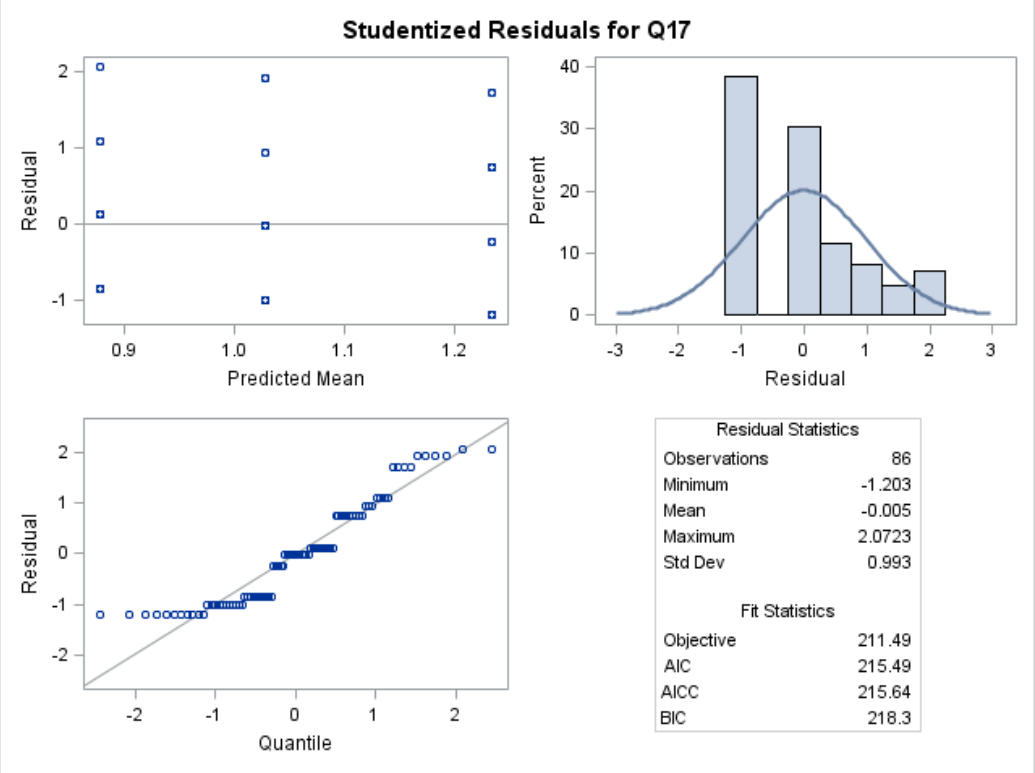
Question 15: Afraid pain would get worse



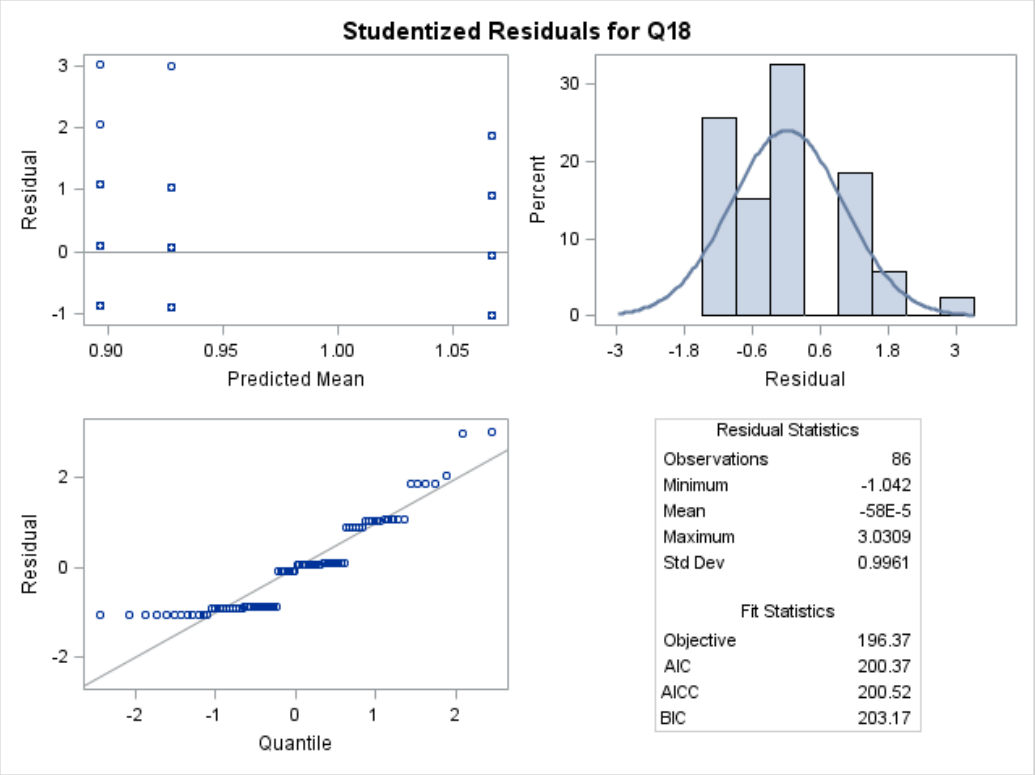
Question 16: Couldn't stand it anymore



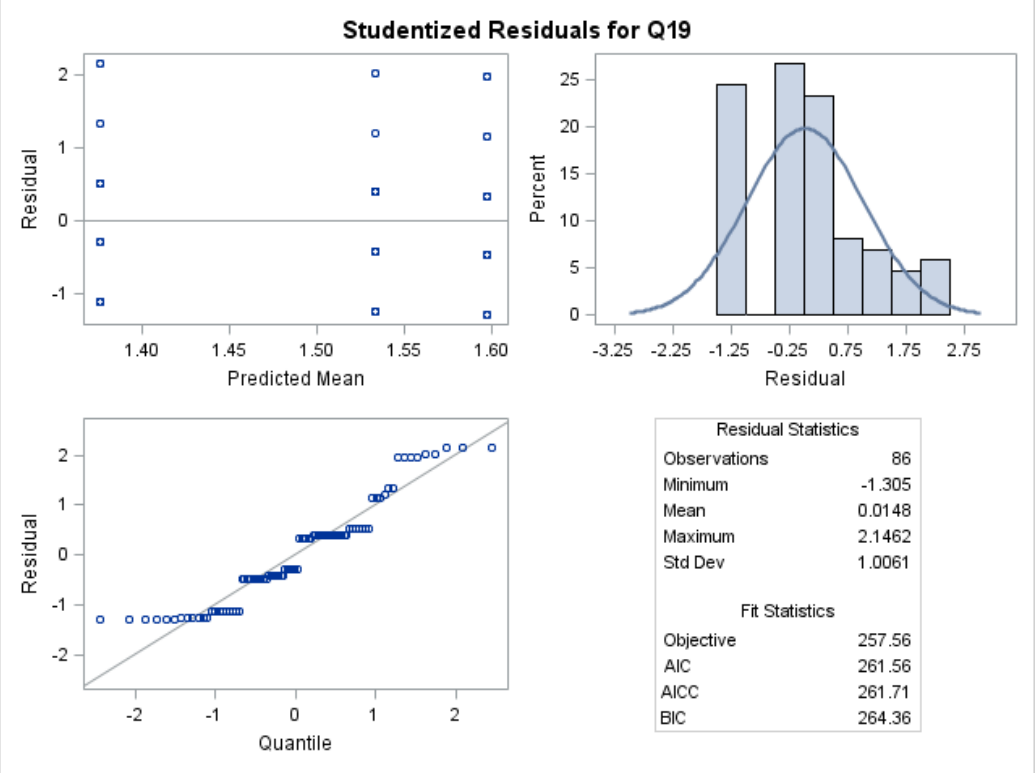
Question 17: Feeling sad



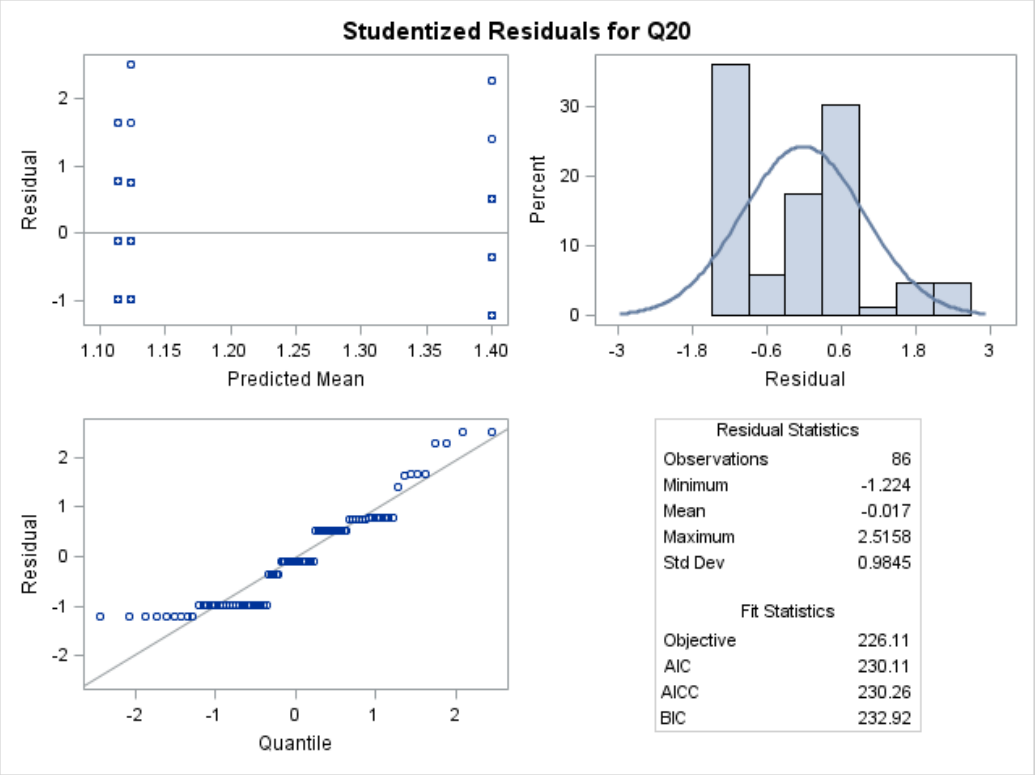
Question 18: Feeling angry



Question 19: Feeling cheerful

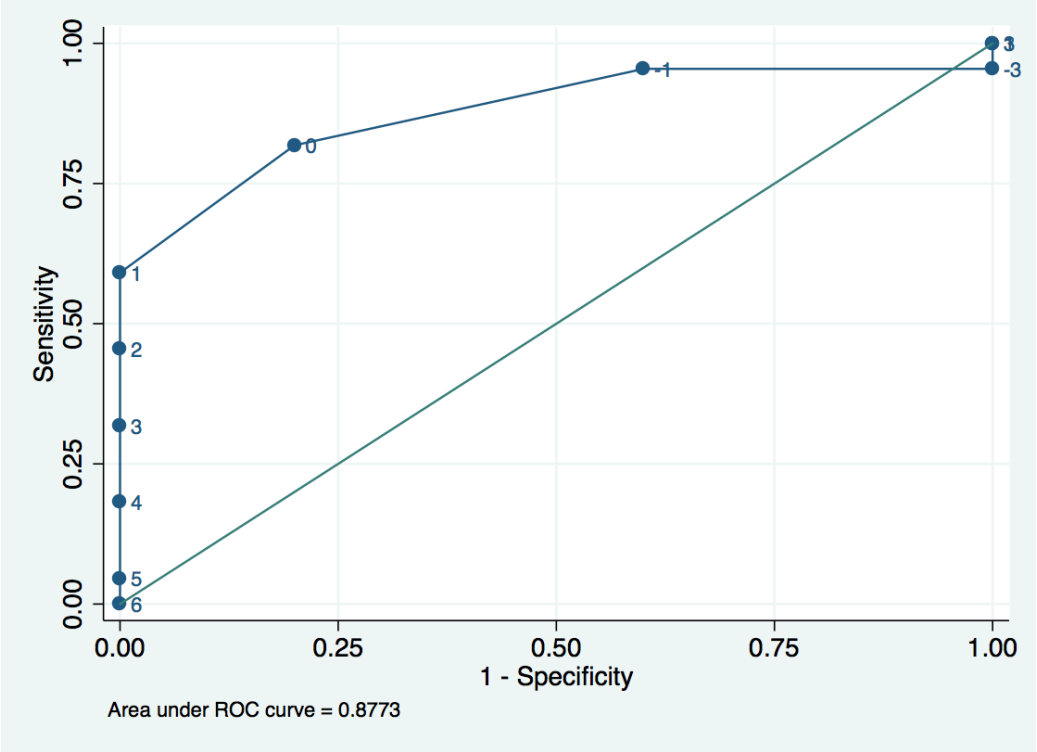


Question 20: Feeling worried

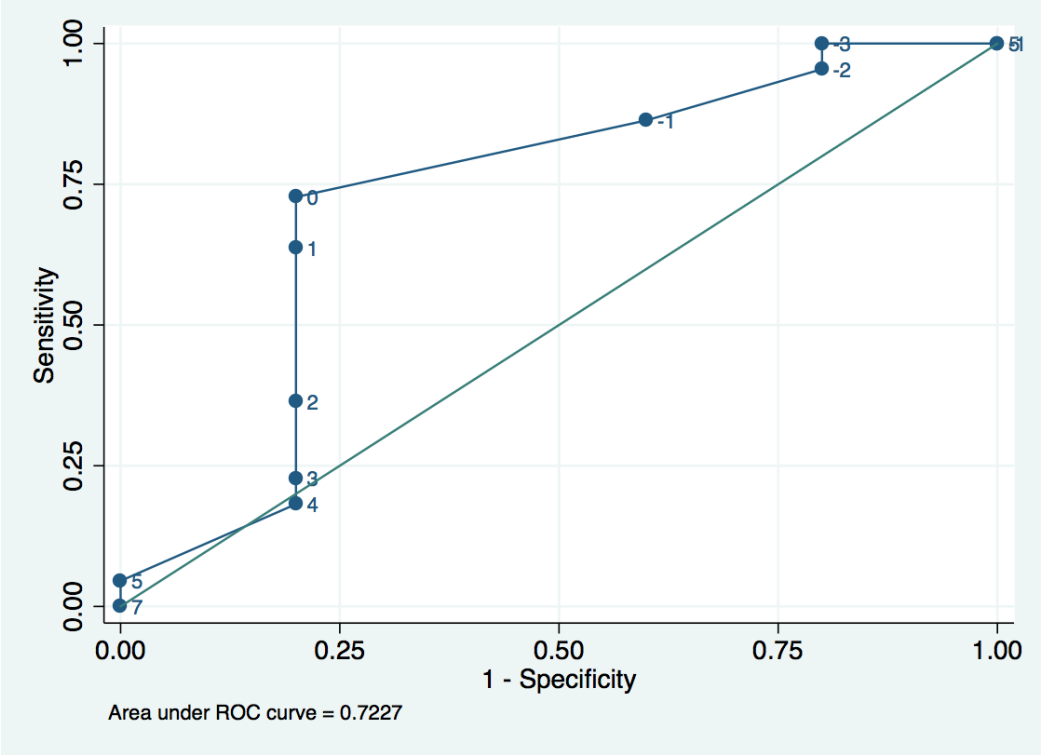


10. ROC Curves for SUPER-KIDZ items: calculated using change scores at 2 weeks post-joint injection

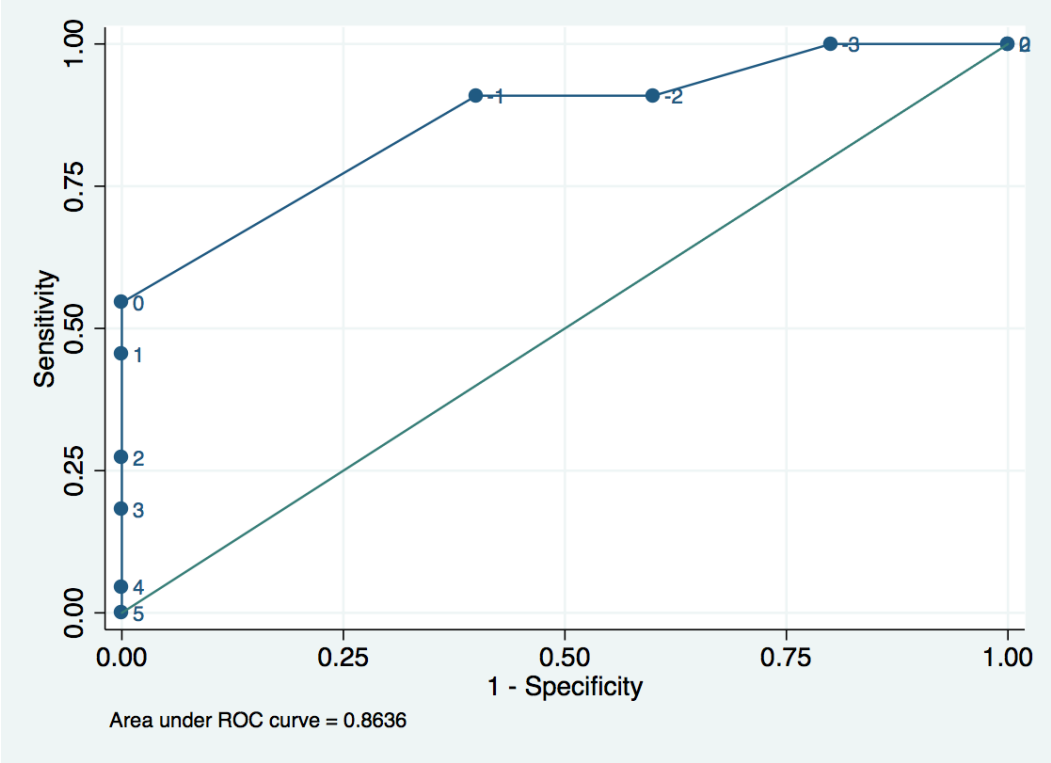
Question 1: Current pain intensity



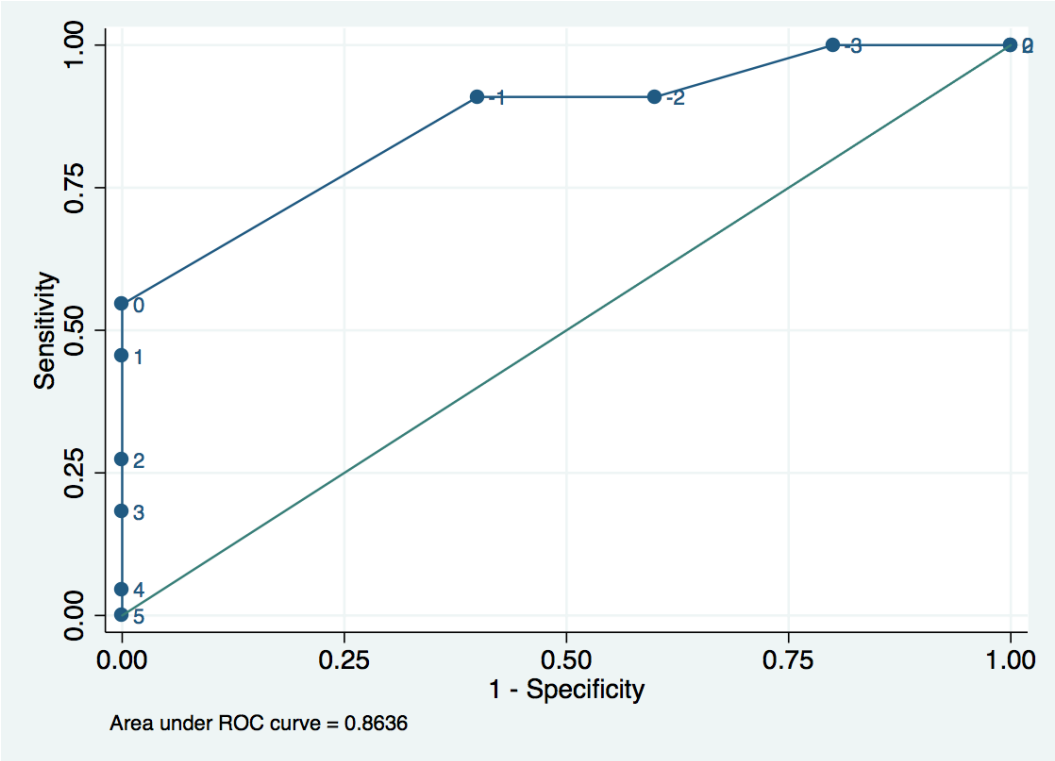
Question 2: Average pain intensity over past week



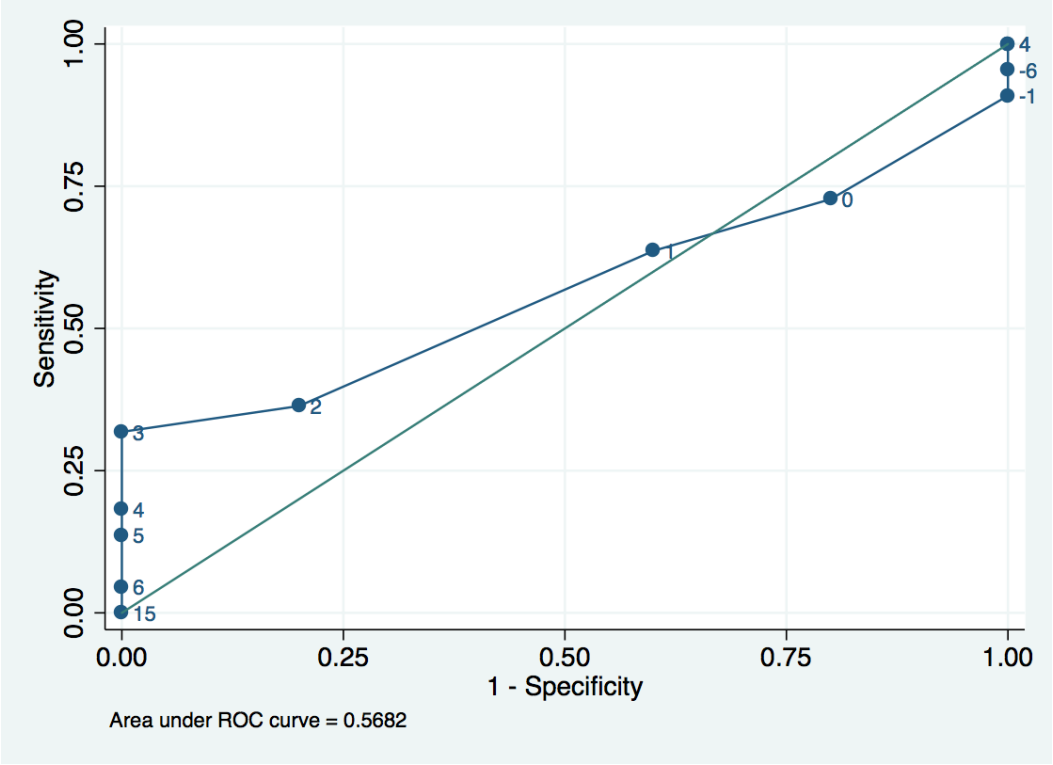
Question 3: Frequency of pain



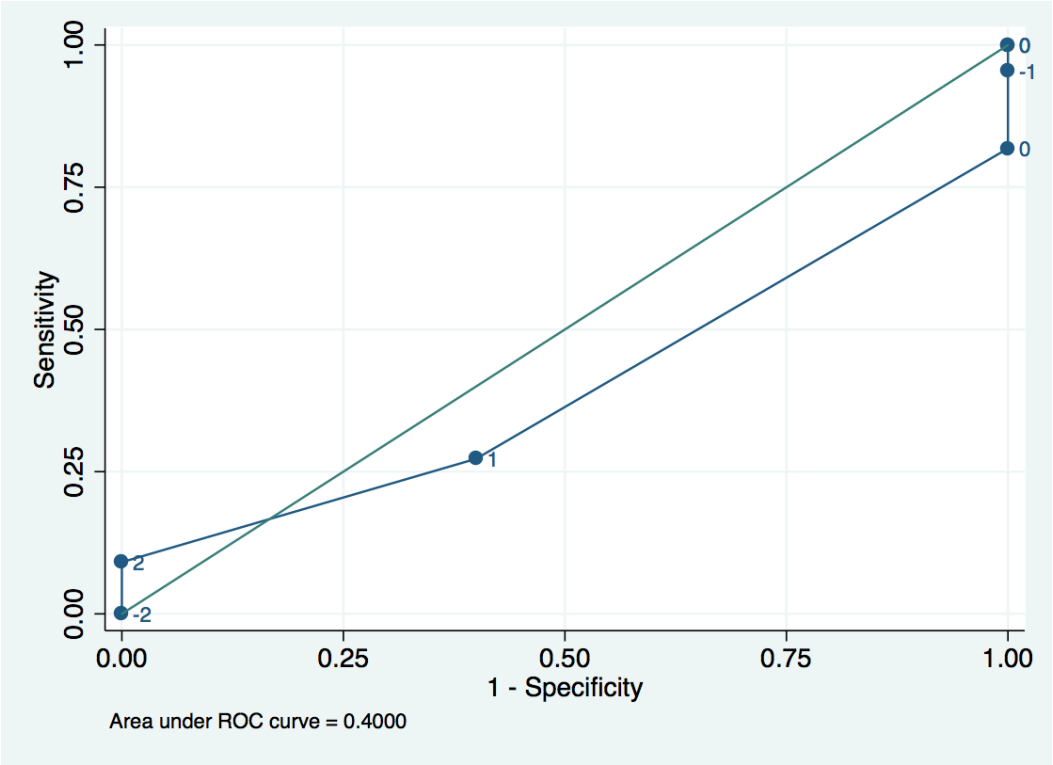
Question 4: Duration of pain



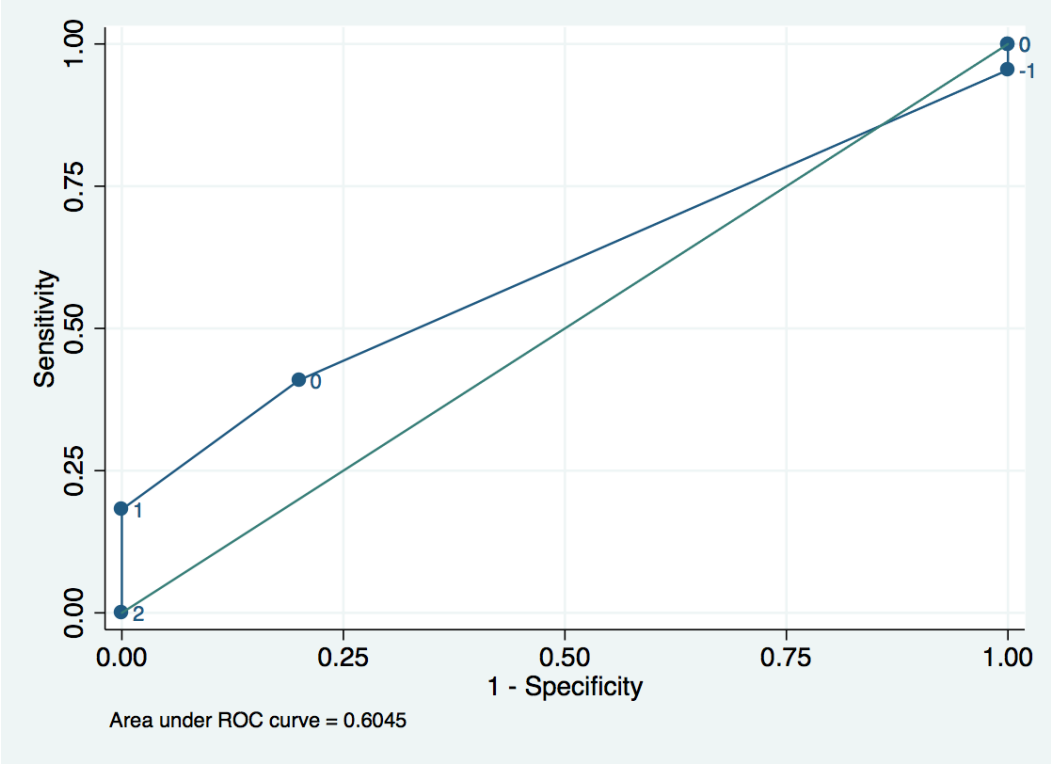
Question 5: Number of painful locations



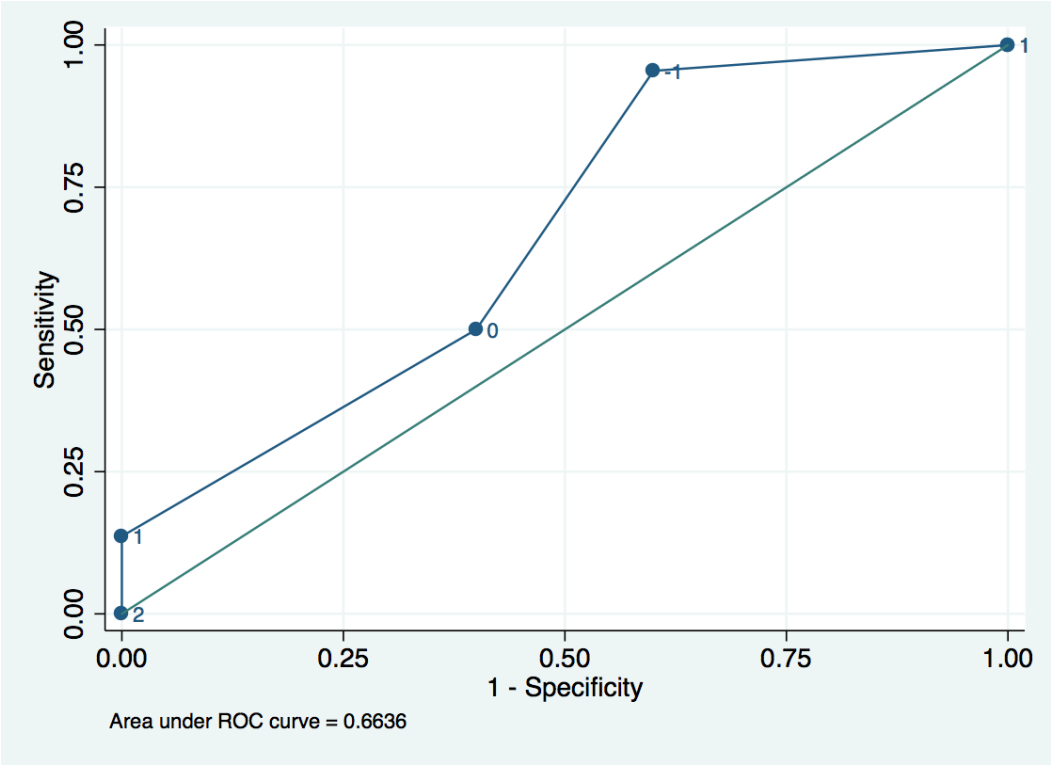
Question 6: Fatigue frequency



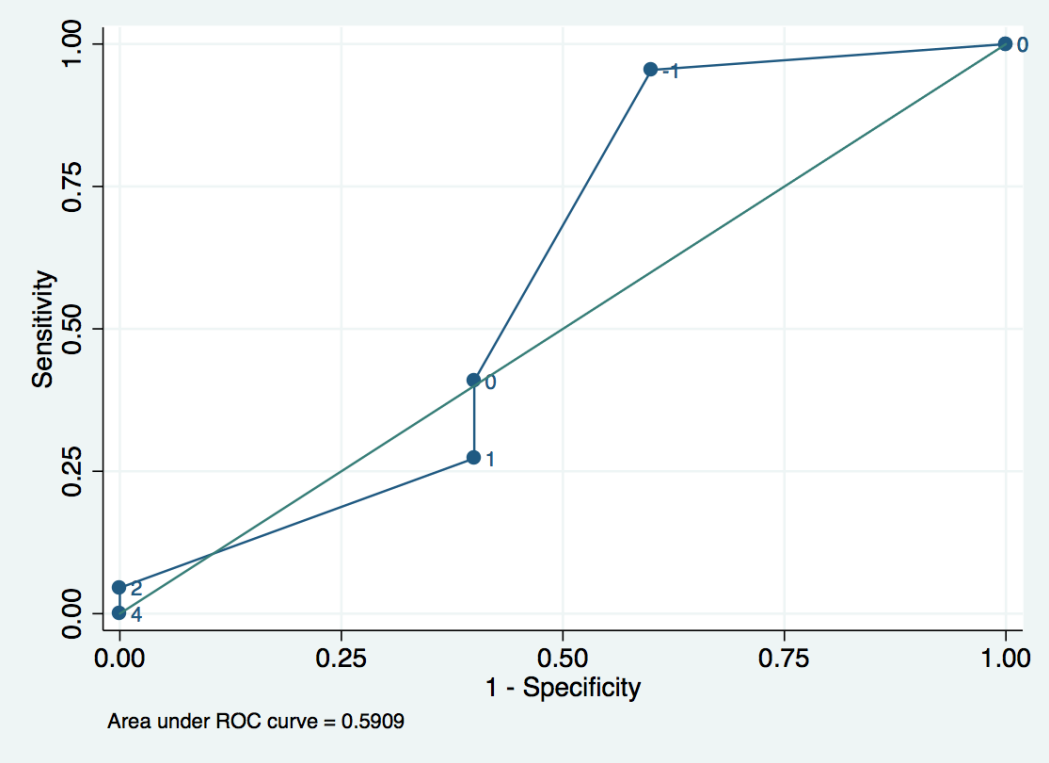
Question 7: Difficulty sleeping



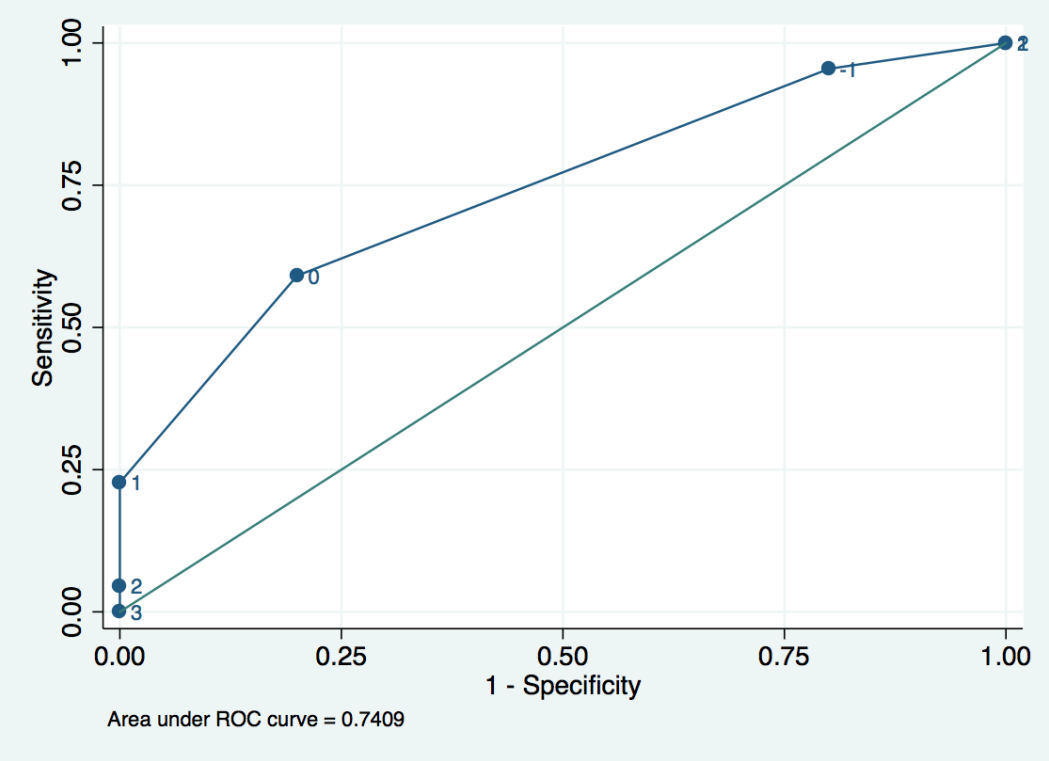
Question 8: Difficulty paying attention



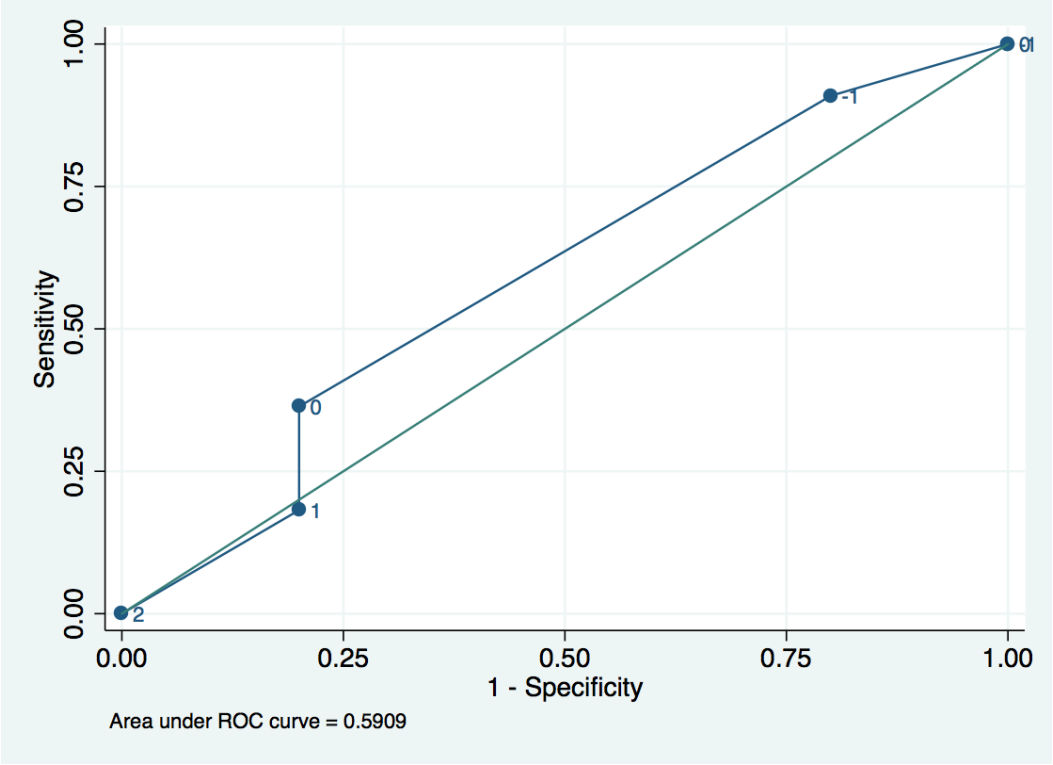
Question 9: Difficulty standing



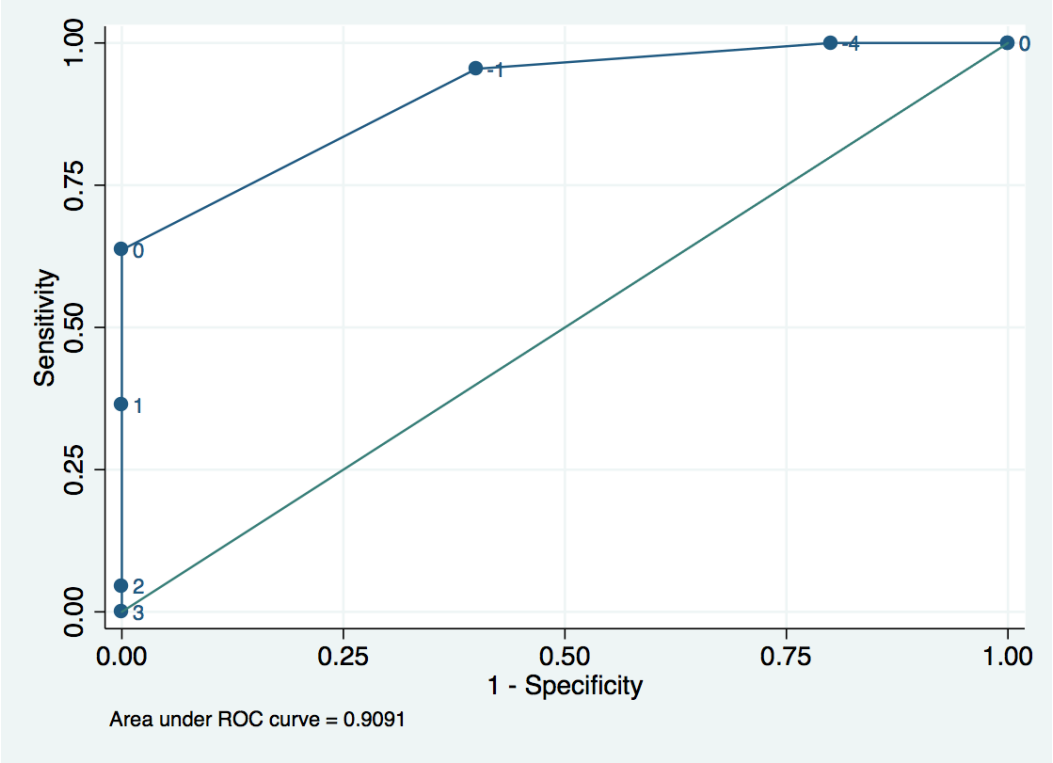
Question 10: Difficulty having fun



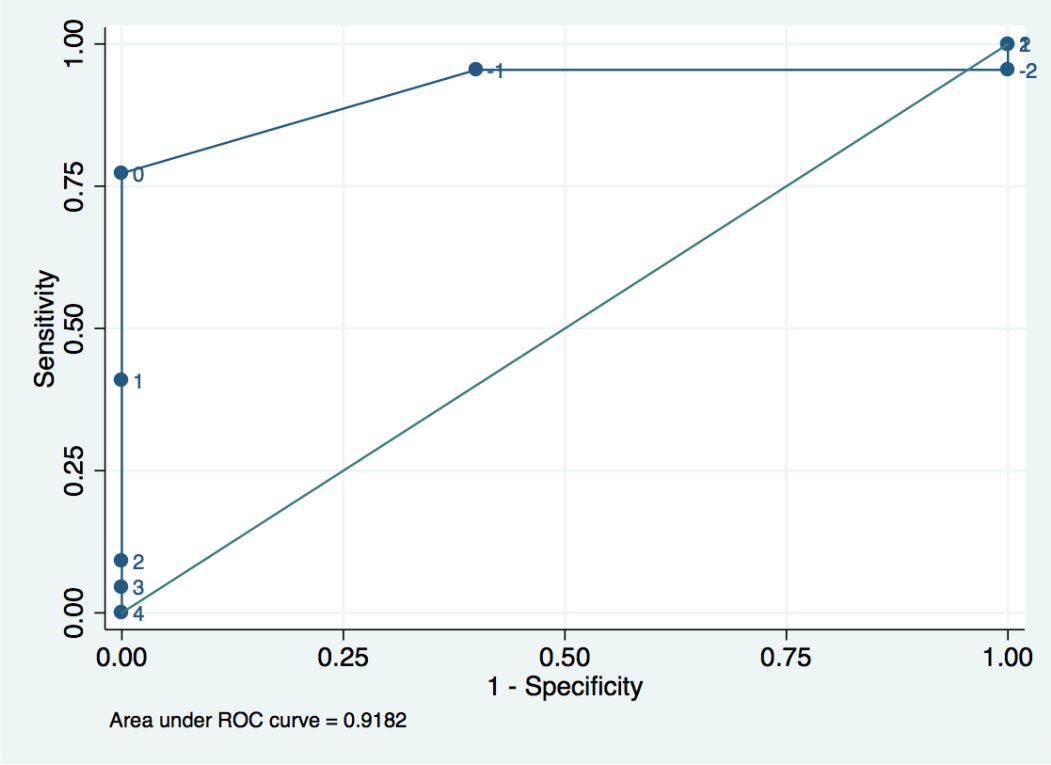
Question 11: Difficulty doing schoolwork



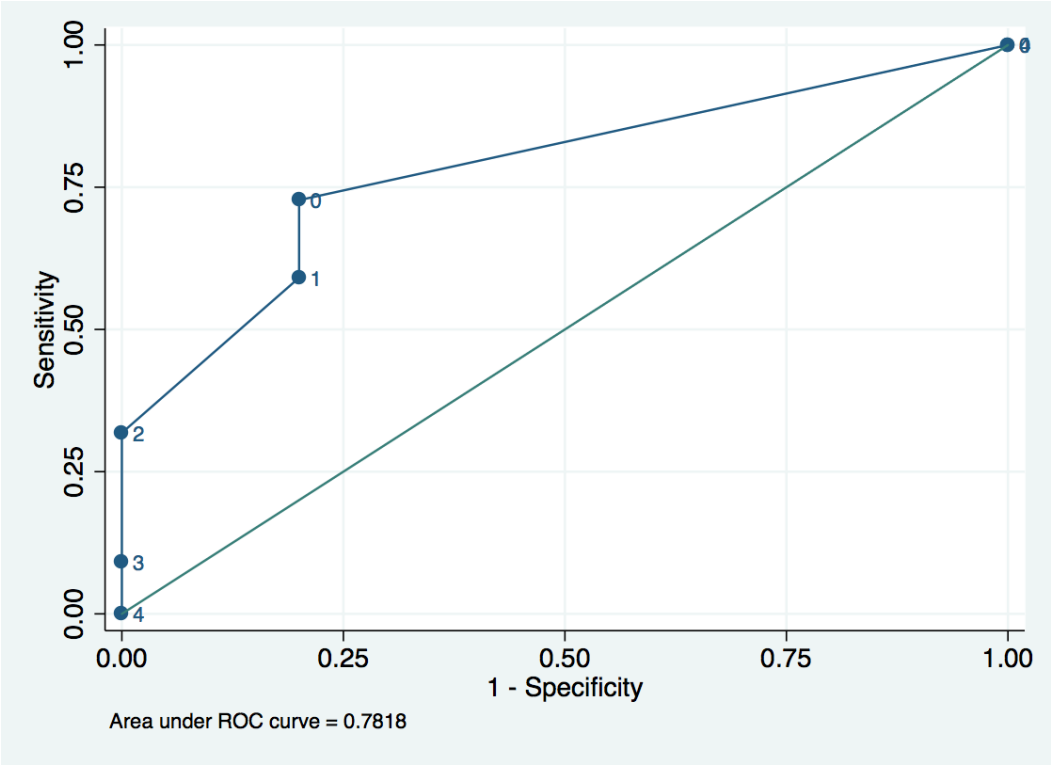
Question 12: Difficulty walking one block



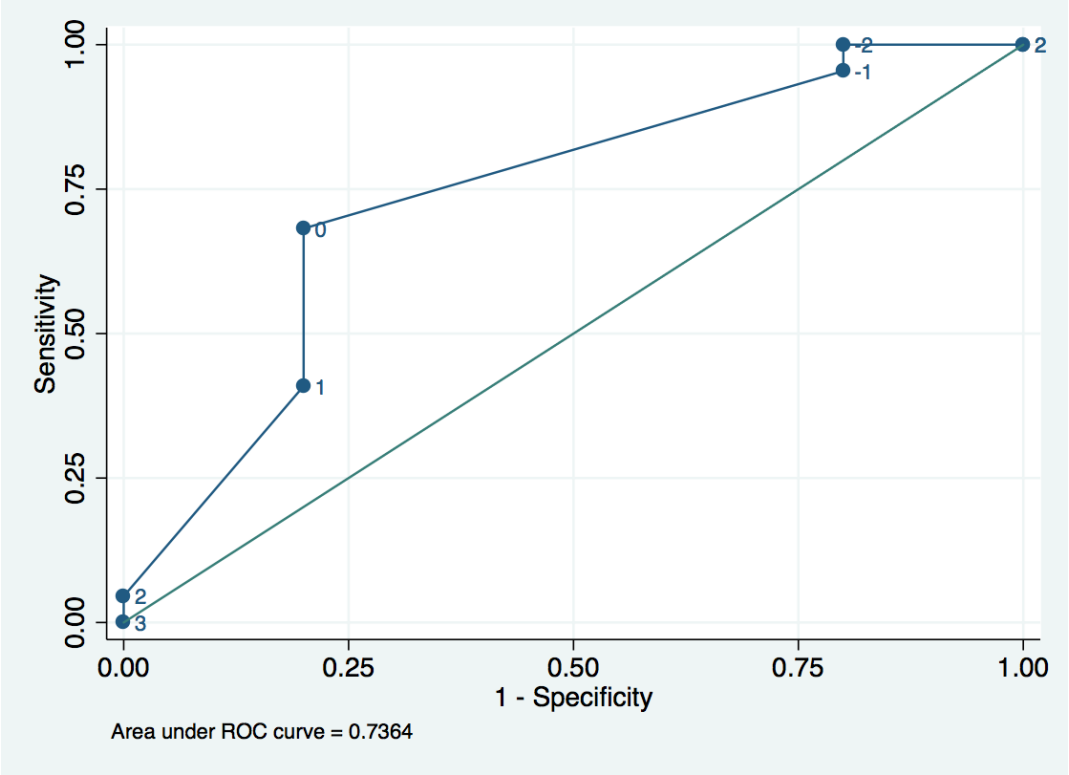
Question 13: Difficulty running



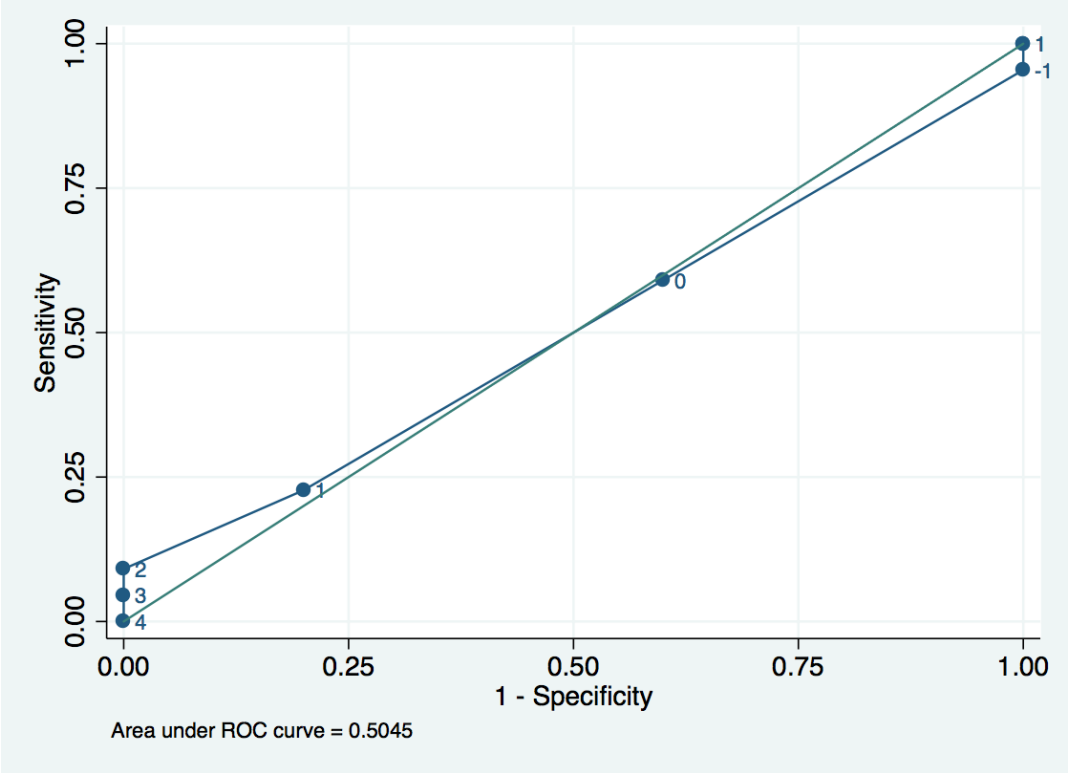
Question 14: Kept thinking how much wanted pain to stop



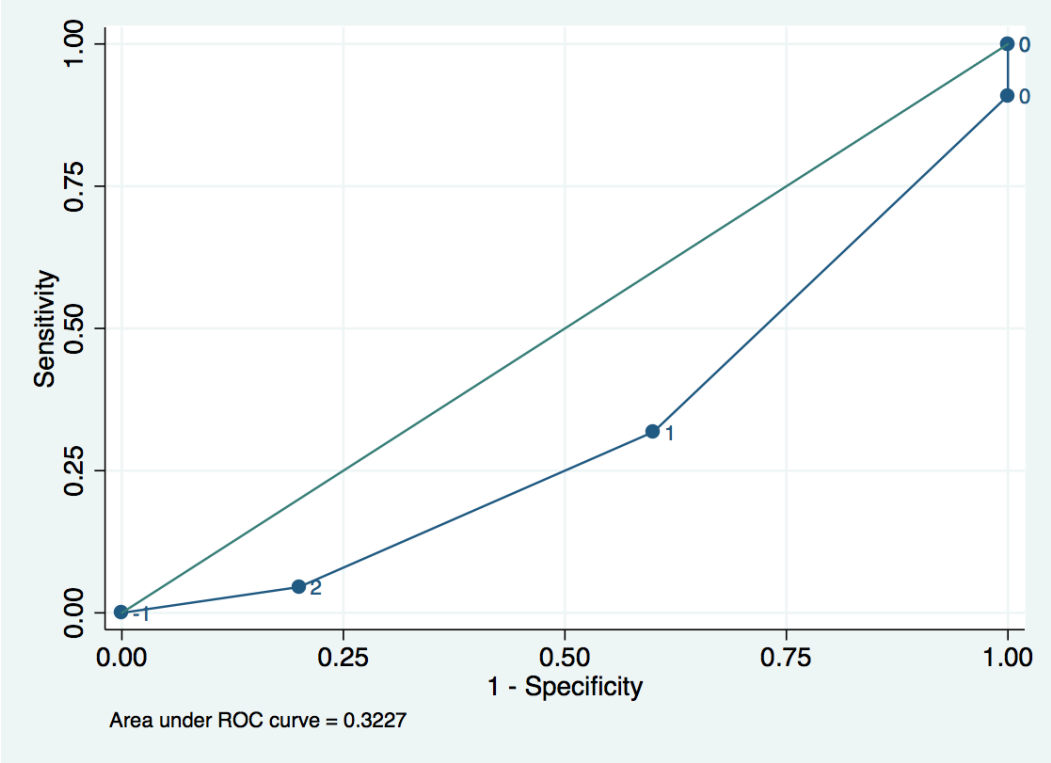
Question 15: Afraid pain would get worse



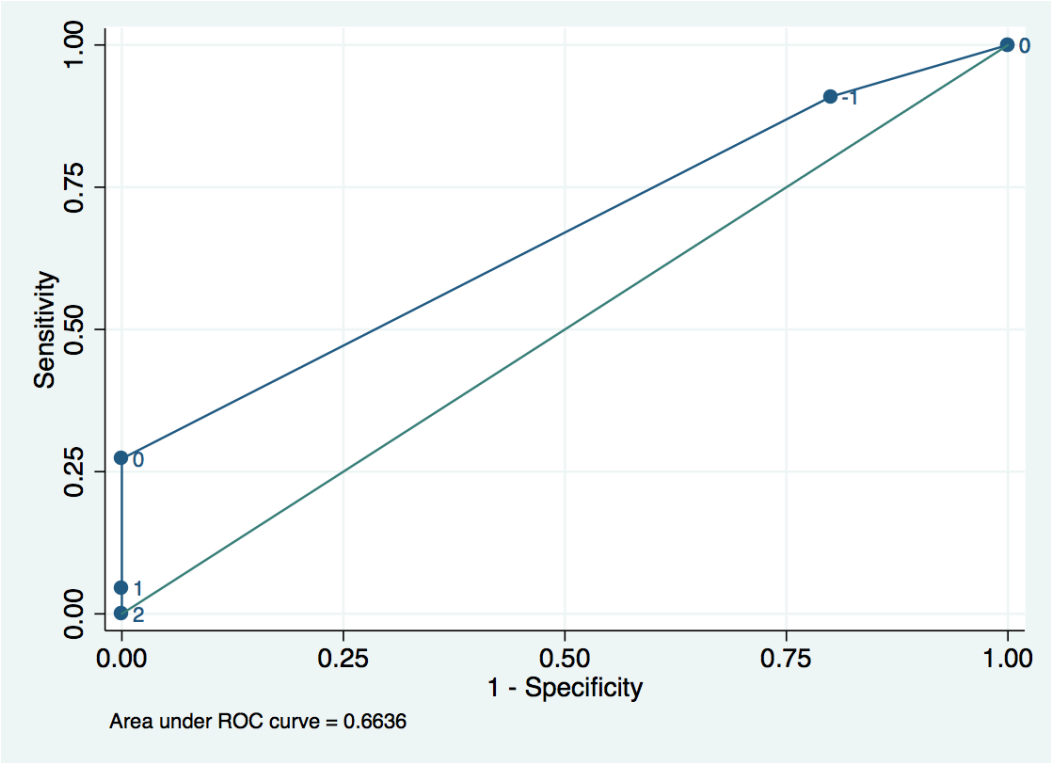
Question 16: Couldn't stand it anymore



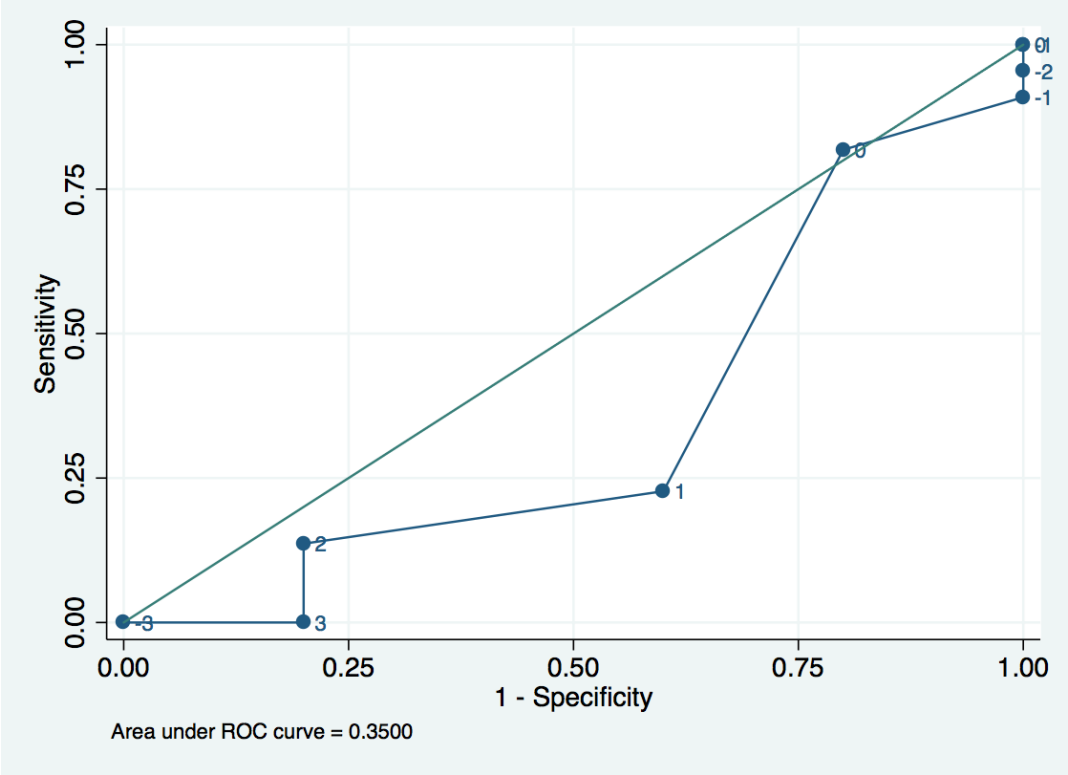
Question 17: Felt sad



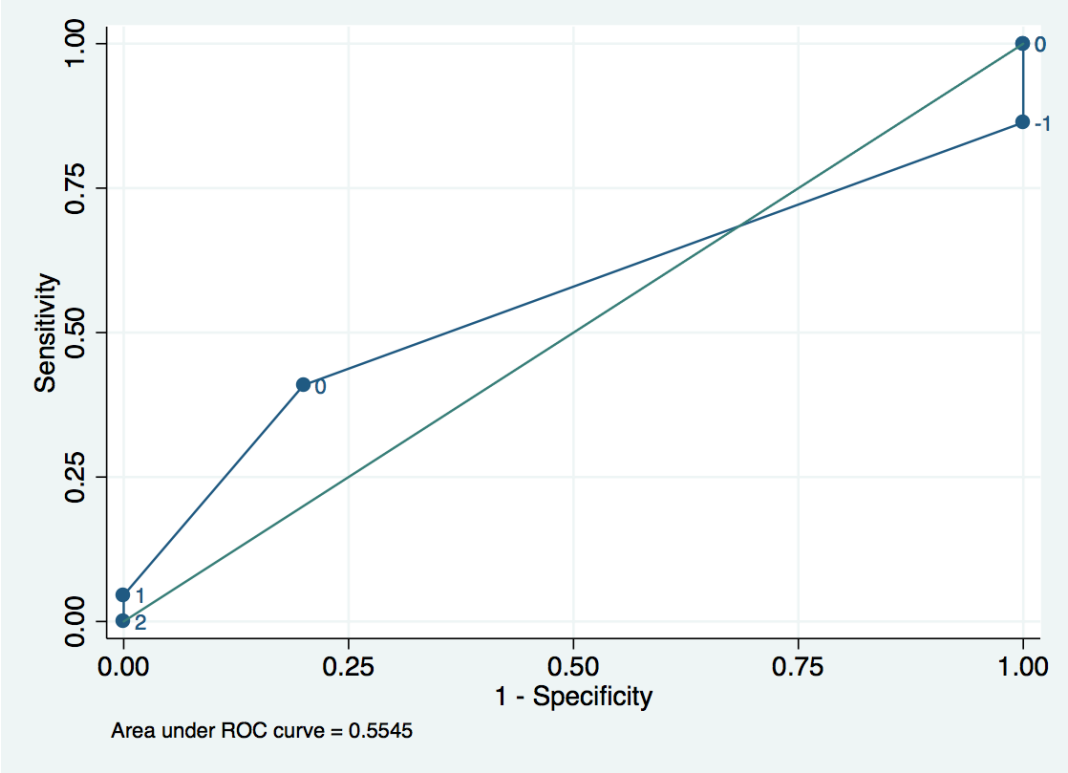
Question 18: Felt Angry



Question 19: Felt cheerful



Question 20: Felt worried



Copyright Acknowledgements

1. Table 3: Definitions of domains of measurement properties. From: Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC: The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010, **63**(7):737-745.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jan 14, 2013

This is a License Agreement between Nadia Luca ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Nadia Luca
Customer address	555 University Ave Toronto, ON M5G2K5
License number	3067691344987
License date	Jan 14, 2013
Licensed content publisher	Elsevier
Licensed content publication	Journal of Clinical Epidemiology
Licensed content title	The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes
Licensed content author	Lidwine B. Mokkink,Caroline B. Terwee,Donald L. Patrick,Jordi Alonso,Paul W. Stratford,Dirk L. Knol,Lex M. Bouter,Henrica C.W. de Vet
Licensed content date	July 2010
Licensed content volume number	63
Licensed content issue number	7
Number of pages	9
Start Page	737
End Page	745
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of	1

https://s100.copyright.com/App/PrintableLicenseFrame.jsp?publisherID=...icationID=12585&rightID=1&typeOfUseID=54&targetPage=printablelicense

RightsLink Printable License

figures/tables/illustrations	
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Reliability and Responsiveness of the Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)
Expected completion date	Jun 2013
Estimated size (number of pages)	120
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.0 USD / 0.0 GBP
Total	0.00 USD

2. Table 4: Summary of intraclass correlation coefficient (ICC) models. From: Weir JP: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005, **19**(1):231-240.

WOLTERS KLUWER HEALTH LICENSE TERMS AND CONDITIONS

Apr 08, 2013

This is a License Agreement between Nadia Luca ("You") and Wolters Kluwer Health ("Wolters Kluwer Health") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Wolters Kluwer Health, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3124260915446
License date	Apr 08, 2013
Licensed content publisher	Wolters Kluwer Health
Licensed content publication	Journal of Strength and Conditioning Research
Licensed content title	Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the Sem
Licensed content author	JOSEPH WEIR
Licensed content date	Jan 1, 2005
Volume Number	19
Issue Number	1
Type of Use	Dissertation/Thesis
Requestor type	Individual
Author of this Wolters Kluwer article	No
Title of your thesis / dissertation	Reliability and Responsiveness of the Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)
Expected completion date	Jun 2013
Estimated size(pages)	120
Billing Type	Invoice
Billing address	555 University Ave Division of Rheumatology Toronto, ON M5G1X8 Canada
Total	0.00 USD
Terms and Conditions	

https://s100.copyright.com/App/PrintableLicenseFrame.jsp?publisherI...a51a7-9987-4b65-b7f7-1edad0caba17%20%20&targetPage=printablelicense

3. Table 5: Axes defining construct of change being measured. From: Beaton DE, Bombardier C, Katz JN, Wright JG: A taxonomy for responsiveness. *J Clin Epidemiol* 2001, 54(12):1204-1217.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Apr 08, 2013

This is a License Agreement between Nadia Luca ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Nadia Luca
Customer address	555 University Ave Toronto, ON M5G1X8
License number	3124260650067
:ense date	Apr 08, 2013
Licensed content publisher	Elsevier
Licensed content publication	Journal of Clinical Epidemiology
Licensed content title	A taxonomy for responsiveness
Licensed content author	D.E Beaton,C Bombardier,J.N Katz,J.G Wright
Licensed content date	December 2001
Licensed content volume number	54
Licensed content issue number	12
Number of pages	14
Start Page	1204
End Page	1217
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
:rmat	both print and electronic

https://s100.copyright.com/App/PrintableLicenseFrame.jsp?publisherI...52667-9c18-4b2b-87e4-2a4333f8684c%20%20&targetPage=printablelicense

Rightslink Printable License

Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Reliability and Responsiveness of the Standardized Universal Pain Evaluations for Rheumatology Providers for Children and Youth (SUPER-KIDZ)
Expected completion date	Jun 2013
Estimated size (number of pages)	120
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.0 USD / 0.0 GBP
Total	0.00 USD