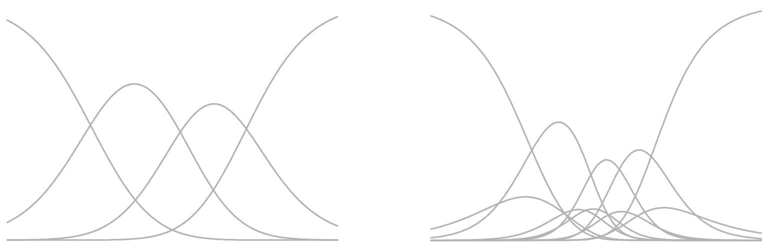
Université catholique de Louvain
Faculté de Médecine
Institut d'Education Physique et de Réadaptation
Unité de Réadaptation et de Médecine Physique

The measure of pain by self-report : use of Rasch analysis

Céline Decruynaere



Thèse présentée en vue de l'obtention du grade de
Docteur en Kinésithérapie et Réadaptation
Orientation: Sciences de la motricité

Promoteurs: Prof. L. Plaghki
Prof. J-L. Thonnard

Mai 2007

Remerciements

Cette thèse représente l'aboutissement de plusieurs années de travail au cours desquelles j'ai pu découvrir le monde de la recherche scientifique. Ces premières lignes sont pour moi l'occasion de prendre le temps de jeter un regard rétrospectif sur ces années et de remercier toutes les personnes qui m'ont entourée.

Je voudrais tout d'abord exprimer ma sincère reconnaissance à mes promoteurs, les Professeurs Léon Plaghki et Jean-Louis Thonnard, pour m'avoir guidée dans la réalisation de ce travail. Bien au-delà de leur rôle de promoteur, ils m'ont emmenée sur les voies de la rigueur scientifique. Je les remercie pour l'enseignement qu'ils m'ont prodigué et l'enthousiasme qu'ils m'ont communiqué.

Le Professeur Marc Crommelinck, les Drs. Massimo Penta, Jacques Grisart et Bernard le Polain de Warroux ont accepté de faire partie du comité d'encadrement de cette thèse. Leurs critiques constructives et leurs conseils m'ont aidé dans mon cheminement. Qu'ils trouvent ici le témoignage de ma sincère gratitude.

Je tiens également à remercier le Dr John Brodersen (Université de Copenhague), les Professeurs Jacques Grégoire, Francis Veyckemans, et Patrick Willems d'avoir participé au jury de cette thèse. Leur critique avertie et leurs conseils m'ont guidée dans la finalisation de ce travail. Qu'ils sachent à quel point je suis honorée de leur présence dans ce jury.

Cette thèse n'aurait pas été possible sans la collaboration enthousiaste des patients du centre de lutte contre la douleur des Cliniques universitaires Saint-Luc et des volontaires sains, adultes et enfants, ayant participé aux expériences de ce travail. Qu'ils trouvent ici le témoignage de ma sincère gratitude. Je tiens également à remercier les personnes qui m'ont permis de recruter et d'évaluer les patients, les Drs Anne Berquin et Etienne Masquelier ainsi que les écoles de Louvain-la-Neuve, Bousval et Dinant qui nous ont permis d'évaluer les enfants.

Mon implication dans le projet de collaboration entre les Professeurs Léon Plaghki et Daniel Le Bars concernant l'exploration psychophysique des nocicepteurs m'a permis de m'ouvrir à d'autres projets de recherche. Les *discussions qui en ont découlé ont été très formatives. Je les en remercie sincèrement.*

Je voudrais aussi remercier tous mes collègues de l'Unité de Réadaptation et de Médecine Physique pour leur soutien et leur amitié. Outre leurs commentaires constructifs lors des staffs et autres discussions plus informelles, la bonne humeur de Carlyne, Laure, Eddy, Yannick, Samar, André, Olivier, Delphine, Thibault, Frédéric, Xavier, Christine, Firas, Gilles, Philippe, Adélaïde et Damienne ont fait du laboratoire de Réadaptation un lieu d'épanouissement professionnel et humain.

Je tiens également à remercier Grégoire Angelou, Vanessa Briart et Noémi Taziaux, mémorants au laboratoire de Réadaptation, pour leur aide efficace dans la collecte des données.

Enfin, pour tout l'amour qu'ils me donnent, mes remerciements s'adressent à Thomas et Irène.

Contents

Summary

The overall aim of this thesis is to demonstrate the usefulness and the efficacy of the Rasch model to provide solutions to a variety of measurement problems in the specific context of pain assessment. As Rasch models rest on the requirement of unidimensionality and as verbal self-report has been described as the most valid measure of the experience of pain (Melzack and Katz, 2006), all chapters of the present work focus on the measurement of a single dimension of the pain experience, its intensity and on self-reported rating scales assessing this dimension.

Chapter 1 presents the basics of the Rasch methodology and describes how one model of the Rasch family of models, the many-facet model, can be applied to the analysis of pain intensity responses following CO_2 laser heat stimulation. Several intensities of stimulation were delivered on the hand of one hundred subjects who had to rate their pain perception on a three-level verbal rating scale. This methodological chapter stressed the ability of the Rasch models to transform the original ordinal scores into linear measures on an interval scale. Advantages and disadvantages of the Rasch methodology are discussed.

Chapter 2 presents a second application of the Rasch methodology: the development and validation of a new measurement tool. In that chapter, an item questionnaire, the Situational Pain Scale (SPS), measuring the mental representation of pain in imaginary painful situations was developed and validated with healthy and chronic pain adults. Chronic pain patients also completed questionnaires assessing clinical and emotional attributes. Responses to the SPS were analyzed according to the rating scale Rasch model to select unidimensional and unbiased items. The final SPS consisted of 18 items with a widespread measurement range, a distribution well-targeted for the sample of healthy subjects and chronic pain patients and a high index of measurement precision. The measure of pain representation was significantly related to present pain intensity and anxiety about pain. Fibromyalgia patients showed significantly higher pain representation measures than all other groups. The psychometric properties of the SPS allows to further test its value to predict behaviors such as chronification of an acute pain condition or coping strategies developed by the chronic pain patient

Chapter 3 detailed a third application of the Rasch models: the investigation and optimization of rating scale functioning. It more specifically examines the question of the optimal number of response levels required to assess the intensity of pain. A first section details the problematic, reviews studies previously conducted and presents the statistical indexes provided by the Rasch methodology for investigating the functioning of rating scales. Two empirical studies follow. They investigate the optimal number of response levels required to assess the intensity of pain by means of the Rasch analysis.

The first study is devoted to the examination of three verbal rating scales (four, five and ten response categories) in healthy adults rating the 18 items of the SPS (developed in chapter 2). Results showed that the 10- and the 5-level verbal rating scales (VRSs) were not effective to assess pain intensity but that the 4-level one was. Recategorization procedures applied on the 10- and the 5-level VRSs revealed that the optimal number of response levels was also around four. In conclusion, this study shows that the human being can only distinguish four response levels when assessing the pain intensity of hypothetical events on a VRS.

The second empirical study is devoted to the examination of faces scales (FSs) in healthy children aged from four to seven years. Results of this second study showed an improvement in the ability to distinguish the response categories with age. The 4-5 year-old children could only distinguish two response categories and the 6-7 year-old children were able to discern the three levels of a three-level faces scale. In conclusion, young children do not distinguish as many faces as proposed by the majority of available faces scales.

Résumé (French summary)

L'objectif de cette thèse est de montrer l'utilité et l'efficacité du modèle de Rasch dans l'apport de solutions à divers problèmes rencontrés lors de la mesure de la douleur. Etant donné que le modèle de Rasch repose sur l'exigence d'unidimensionnalité et que l'auto-évaluation par le sujet a été décrite comme étant la mesure la plus valide de l'expérience de la douleur (Melzack and Katz, 2006), tous les chapitres de la thèse se sont focalisés sur la mesure d'une seule dimension de la douleur, son intensité, et sur les échelles d'auto-évaluation mesurant cette dimension.

Le **chapitre 1** présente les fondements de la méthodologie Rasch et décrit comment un des modèles de Rasch, le modèle multi-facettes, peut être appliqué à l'analyse de scores d'intensité de douleur fournis en réponse à des stimulations de chaleur délivrées à l'aide d'un laser CO_2. Plusieurs intensités de stimulation ont été délivrées sur la main de cent sujets sains qui devaient évaluer leur perception de douleur sur une échelle verbale simple à trois niveaux de réponse. Ce chapitre méthodologique met l'accent sur la capacité du modèle de Rasch à transformer les scores originaux ordinaux en mesures linéaires sur une échelle d'intervalles. Les avantages et inconvénients de la méthodologie Rasch y sont discutés.

Le **chapitre 2** présente une seconde application de la méthodologie Rasch: le développement et la validation d'un nouvel instrument de mesure. Dans ce chapitre, un questionnaire, le Situational Pain Scale (SPS), mesurant la représentation mentale de la douleur lors de situations de douleur imaginaires est développé et validé dans une population d'adultes sains et de patients douloureux chroniques. Les patients douloureux chroniques ont également rempli une batterie de questionnaires évaluant diverses variables cliniques et émotionnelles. Les réponses au SPS ont été analysées à l'aide du modèle *rating scale* afin de sélectionner les items satisfaisant au critère d'unidimensionnalité et d'invariance de l'échelle. L'échelle finale comprend 18 items présentant une large étendue de mesure, une distribution bien ciblée par rapport à l'échantillon de sujets sains et de patients douloureux chroniques et une précision de mesure élevée. La mesure de la représentation de la douleur est significativement corrélée avec l'intensité de la douleur au moment de l'évaluation ainsi qu'avec l'anxiété spécifiquement liée à la

douleur. Les patients fibromyalgiques ont une représentation de la douleur plus élevée que tous les autres groupes. Les qualités psychométriques de l'échelle permettent, dans le futur, de tester sa capacité à prédire certains comportements comme par exemple le risque de chronification d'un état de douleur aiguë ou encore les stratégies développées par le patient pour faire face à son état de douleur chronique.

Le **chapitre 3** détaille une troisième application des modèles de Rasch : l'investigation et l'optimisation du fonctionnement des échelles de réponse. Ce chapitre examine plus spécifiquement la question du nombre optimal de niveaux de réponse nécessaire pour évaluer l'intensité de la douleur. Une première section introduit la problématique, présente une revue de la littérature sur la question et détaille les indices statistiques fournis par la méthodologie Rasch pour étudier le fonctionnement des échelles de réponse. Deux études empiriques suivent. Elles étudient le nombre optimal de niveaux de réponse nécessaire pour l'évaluation de la douleur.

La première étude est consacrée à l'investigation de trois échelles verbales simples (quatre, cinq et dix catégories de réponse) chez des adultes sains évaluant les 18 items du SPS dont le développement a été présenté au chapitre 2. Les résultats montrent que les échelles verbales simples (EVSs) à dix et cinq niveaux ne sont pas efficaces pour évaluer l'intensité de la douleur alors que l'échelle à quatre niveaux l'est. Les procédures de recatégorisation appliquées sur les EVSs à dix et cinq niveaux révèlent que le nombre optimal de niveaux de réponse est de quatre. En conclusion, cette étude montre que l'être humain ne peut distinguer que quatre niveaux de réponse lorsqu'il évalue des situations imaginaires sur une EVS.

La seconde étude empirique est consacrée à l'investigation des échelles visages chez des enfants sains âgés de quatre à sept ans. Les résultats de cette seconde étude montrent une amélioration dans la capacité à distinguer les catégories de réponse avec l'âge. Les enfants de quatre et cinq ans ne peuvent distinguer que deux catégories de réponse alors que les enfants de six et sept ans peuvent distinguer les trois niveaux d'une échelle à trois visages. En conclusion, les jeunes enfants ne peuvent pas distinguer autant de visages que proposés par la majorité des échelles visages publiées.

Introduction

The popular definition of pain is that pain is 'what hurts' (Sternbach, 1989). This definition highlights that we all recognize its existence and know what we mean by it. The International Association for the Study of Pain defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" (IASP, 1979). This latter definition acknowledges the importance of affective-motivational as well as sensory-discriminative factors and underscores the inherent subjectivity of pain. The sensory-discriminative dimension refers to the sensory qualities of pain, such as the intensity, location, duration and quality. The affective-motivational dimension reflects the emotional and aversive aspects of pain. In addition, Melzack and Casey (1968) define a cognitive-evaluative dimension reflecting the contextual meaning, past experiences and expectations. These three dimensions interact with one another to generate the complex perception and behavior that characterize pain. The quantification of pain is crucial in order to assess the needs of patients, to follow pain over time, to evaluate the effectiveness of treatments and, if necessary to adjust therapy. However, pain cannot be directly observed since it is a latent variable, that is a variable hidden within the person such as intelligence or anxiety. It can only be inferred from responses of the subject (Turk, 1989). There are three major classes of pain responses: motor activity, physiological variables and self-report.

Motor activity

It is well-known that pain elicits stereotypical behaviours. These include reflexive withdrawal, limping, rubbing, moaning, facial expressions and vocalization. A wide variety of observational techniques has been developed to assess these behaviours (Keefe et al., 1992). These techniques are especially useful for assessing pain in infants and preverbal children who lack language skills or in adults with mental, psychological, attention or communicative disorders. However, there are a number of important issues that arise when trying to incorporate behavioural observation methods into clinical practice settings. These issues comprise the time spent observing behaviour and the need for observer training. Additionally, while experience and behaviour are usually highly correlated, they are far from identical. For example, discordance is often observed between healthcare providers and patients ratings of patient pain (Choinière et al., 1990; Coulthard et al., 2000; Guru and Dubinsky, 2000). Finally, behavioural responses may also be indicative of anxiety or stress instead of pain; questioning the validity of these techniques for measuring pain. For these reasons, behavioural assessments of pain should not replace verbal self-report if patients are able to rate their pain.

Physiological variables

Physiological changes often accompany the experience of pain. Physiological correlates of pain that are frequently measured include heart rate, blood pressure, skin conductance, electromyographic activity and cortical evoked potentials (Melzack and Katz, 2006). Although such physiological measures can be useful, great caution should be exercised. Some investigators have placed great faith in physiological measures because they appear more objective than self-reports as they use physical units of measurement to quantify pain. However, objective does not mean more valid. Responses observed in these physiological variables are only correlates of pain and are not specific to the experience of pain. A great variety of situations of general arousal and stress will evoke these physiological responses. Moreover, physiological responses habituate over time despite the persistence of pain (Turk, 1989).

Self-report

Because pain is subjective, the patient's self-report has been described as the most valid measure of a patient's experience of pain (Melzack and Katz, 2006). The present manuscript will entirely focus on this category of pain responses. Most commonly used self-report rating scales are the visual analogue scale (VAS), the numerical rating scale (NRS) and the verbal rating scale (VRS). These scales are considered the instruments of choice when a unidimensional measure of pain is required. Very often, the sensory-discriminative dimension of pain is assessed through these methods. The sensory-discriminative dimension refers to the sensory qualities of pain, such as intensity. Nevertheless, the affective dimension of pain is also assessed with these self-report rating scales. The affective dimension of pain reflects the emotional and aversive aspects of pain. The verbal anchor points on a VAS can be modified to delineate the different dimensions of pain (For example, 'no pain sensation' and 'most intense pain sensation imaginable' for the sensory-discriminative dimension *versus* ‘not at all unpleasant’ and ‘most unpleasant imaginable’ for the affective dimension).

These scales are easy to design, to administer and to score, and provide quite valuable information. However, traditional psychometric analyses conducted on these scales have many well-recognized limitations. Two particularly relevant limitations for clinical practice are underlined here. The first concerns the strongly ordinal properties of scores provided by these scales (Chapman et al., 1985; Merbitz et al., 1989). A numerical value is generally assigned to each response category and when several questions about pain (or items) are grouped together to form an instrument, frequently, arithmetic operations are performed on scores. For instance, the raw scores of each item are summed generating a total score that quantifies pain. However, if categories imply rank-ordering, intervals between their boundaries are not necessarily equal. The size of each category is difficult to specify. Consequently, such operations are not valid. Arithmetic operations require scales with known and equal intervals. Such scales are defined as interval scales according to the classification of measurement scales established by Stevens (1946). Figure 1 illustrates the difference between an ordinal and an interval temperature scale. It can be seen that the distance between a score of 1 (cold) and 2 (mild) on the ordinal scale (bottom scale) is not the same as the distance between the distance between for example a score of 4 (hot) and 5 (scorching). On the other hand, distances between each pair of adjacent graduation expressed in degrees Celsius on the interval scale are always the same. The second limitation is the dependency of the patient’s score to the particular scale used to assess him. In other words, a subject’s level of pain will depend on the items used. Scales including less or more painful items will lead to different pain scores while the patient’s level of pain remains the same.

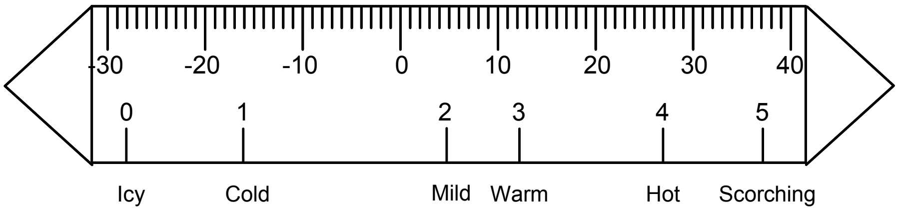


Figure 1. A ruler with an ordinal scale at the bottom and an interval scale at the top. Both scales express the temperature. The ordinal scale comprises six verbal categories and an score is allocated to each category. The interval scale is a temperature measurement scale expressed in degrees Celsius.

Item Response Theory

Because pain is a latent trait, psychometric methods assessing latent traits also called Item Response Theory (IRT) may be particularly useful in pain measurement (Hambleton et al., 1991; Laveault and Grégoire, 2000; Bertrand and Blais, 2004). Interest has focused on these methods because of their potential to overcome the limitations of traditional psychometric methods applied on ordinal scales mentioned just here above (Turk et al., 2006). The measure of the pain the respondent feels is estimated through his responses to a set of calibrated items. Several responses are needed to obtain a stable estimate as the probability of choosing a given response (for instance, not painful) to a given item is not perfect. The painfulness conveyed by the choice of either word may not exactly match the feeling of the respondent for this item. But, when this procedure is repeated for several items, a good estimate of the respondent's level of pain is achieved. In essence, IRT methods model the probability of an individual's response to an item. The relationship between the probability of success[a] to an item and the latent trait (e.g., the pain of the subject) is described by a function called item characteristic curve (ICC) that takes an S-shape (Figure 2).

[a] 'Success' in the context of the particular case of the pain variable could be a response 'painful' in a dichotomous response choice 'not painful' / 'painful' scored 0 / 1.

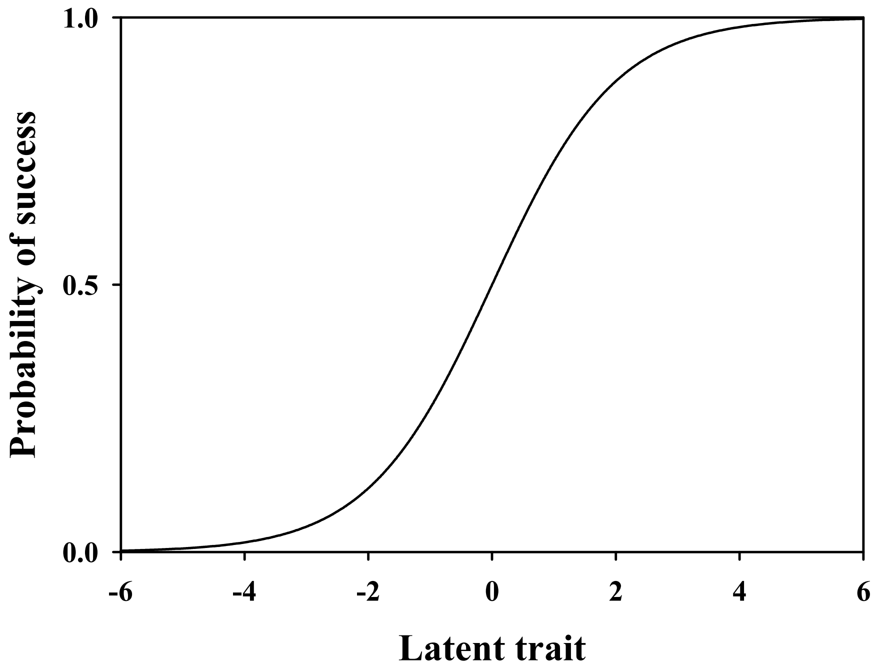


Figure 2. An item characteristic curve showing the relationship between location on the latent trait and probability of answering the item correctly.

The three most popular IRT models are the one-, two-, and three-parameter logistic models, so named according to the number of parameters describing the item. A logistic function has been preferred over a cumulative normal function because its formulation is much easier to deal with, more mathematically tractable and its graphical representation roughly the same (Andrich, 1978a; Massof, 2002; Bertrand and Blais, 2004).

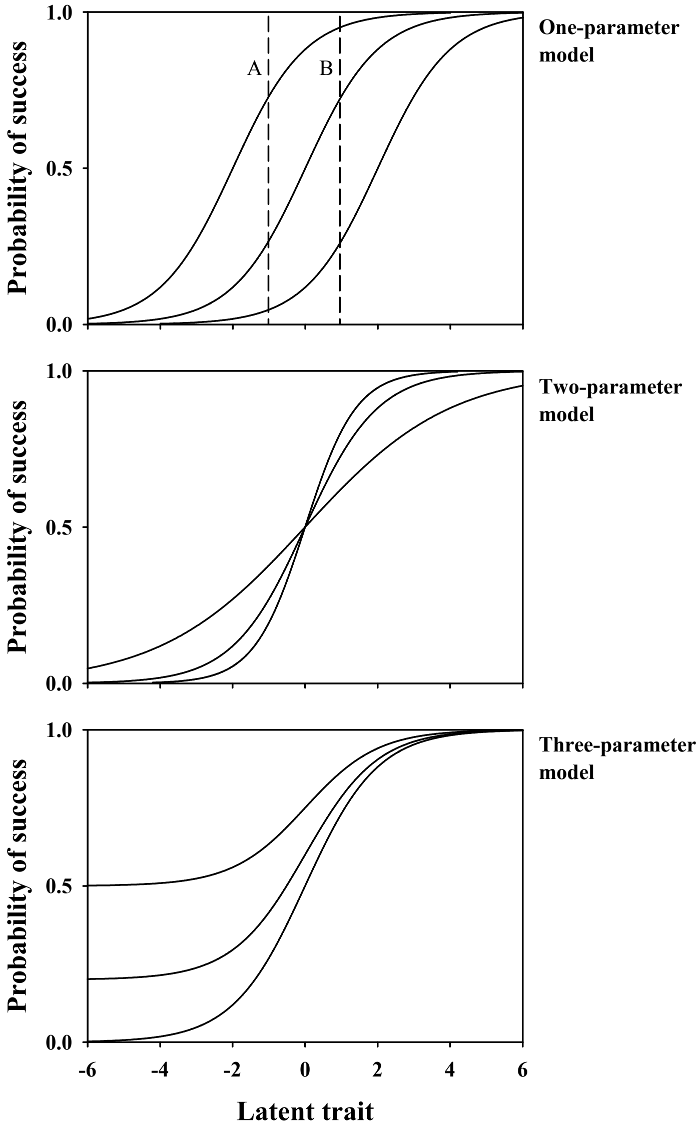


Figure 3. Top panel: item characteristic curves for three items according to a one-parameter model. More difficult items are shifted to the right. The vertical discontinuous lines represent the ability of two subjects (A: less able; and B: more able). For both subjects, the item hierarchy remains unchanged. Middle panel: item characteristic curves for three items according to a two-parameter model. The difficulty of all three items is the same. The slope of the curve reflects the item discrimination. The item with the steeper curve is more discriminating in the vicinity of the inflexion point (the difficulty of the item) but less discriminating at both ends of the scale. Bottom panel: item characteristic curves for three items according to a three-parameter model, with a different pseudo-chance-level parameter. The nonzero lower asymptote corresponds to the probability that a subject can get the item correct by guessing. The probability of answering correctly one item mainly varies for subjects with a low ability.

For the one-parameter logistic model, the parameter that varies across items is the difficulty.[b] Item characteristic curves differ only by their location along the scale (Figure 3, top panel). The curves do not cross; they run parallel. By convention, the difficulty of the item is the value of the latent trait for which the probability of success is 0.5. Several one-parameter models exist but, most of the time, only one is used, the Rasch model (Bertrand and Blais, 2004; Rasch, 1960). For the two-parameter logistic model, an additional parameter, called the item discrimination parameter, may vary across items. The discrimination parameter characterizes the slope of the curve at the inflexion point. Three ICCs for a two-parameter model are shown in the middle panel of Figure 3. All items have the same difficulty. The curves differ according to their slope and consequently cross. In the three-parameter model, a third parameter called the *pseudo-chance-level* is added providing a possible nonzero lower asymptote for the ICC. This model is used when guessing may influence the response of subjects with low ability as is the case in multiple choice items for which even less able subjects have a certain probability of answering the item correctly. Three-parameter logistic ICCs are displayed in the bottom panel of Figure 3. The probability of success mainly varies for less able subjects. The second and the third parameters in the two- and the three-parameter IRT models are introduced and manipulated to maximize the fit of the model to the data.

The Rasch model

A one-parameter model, the Rasch model, is broadly and successfully used in practical health care applications (Conrad and Smith, 2003). Several factors explain this intense popularity.

First, because of its simplicity, smaller samples are required for estimation of item difficulty parameters compared with other IRT models that use item discrimination and guessing parameters (Thorndike, 1999; McHorney and Monahan, 2004).

[b] 'Difficulty' in the context of the particular case of the pain variable could be relabeled 'painfulness'.

Second, the availability of user-friendly software, such as Winsteps (Linacre and Wright, 1998) and RUMM (Andrich et al., 2003) is a key-aspect that enables clinicians and researchers to apply the Rasch methodology.

The third and most important factor, the Rasch model, developed independently but concomitantly to IRT is not simply algebraically a special case of more general IRT models but differs from IRT at the most basic level. The Rasch model is the only one among IRT models to generate fundamental measurements of the kind found in the physical sciences. This arises from the fact that the Rasch model is the only one to provide a mathematical formalization of "specific objectivity" which turns out to be an operational criterion for fundamental measurement. The property of "specific objectivity" implies that comparisons between individuals are independent of which particular items within the class considered have been used. Symmetrically, it is possible to compare items belonging to the same class – measuring the same thing - independently of which particular individuals within a class considered answered them (Rasch, 1960; Andrich, 2004). This property is linked to the fact that ICCs run parallel and do not cross in the Rasch model (Figure 2, top panel). This prevents the hierarchy of difficulty of the items from changing depending on the ability of the subjects (Tesio, 2003). Other IRT models including two or three parameters make the curves cross and the hierarchy of items change. Technically, this property of specific objectivity is related to the specific logistic formulation of the Rasch model that transforms the total score of each subject into a measure of ability independent of the relative difficulty of the particular items considered and, similarly, transforms the total score of each item into a measure of difficulty independent of the ability of the particular sample considered (Rasch, 1960). In other words, a person's total score is sufficient for estimating its 'Rasch' ability and an item total score is sufficient for estimating its 'Rasch' difficulty. No information is contained in the pattern of responses.

The Rasch model requires that only one trait is measured by the set of items. This is called the requirement of unidimensionality (Rasch, 1960; Bond and Fox, 2001; Hobart, 2002; Massof, 2002; Andrich, 2004). The ideal in terms of evidence of unidimensionality is the Guttman scale (Guttman, 1950). A Guttman scale is represented in Table 1. This data matrix shows the response pattern produced by twelve persons answering twelve items. Persons, illustrated by letter a to l, are presented from top to bottom in order of decreasing ability; items, illustrated by numbers 1 to 12 are presented from left to right in order of increasing difficulty. Guttman scales imply a perfect scoring pattern in the sense that persons must answer correctly to items in order of difficulty until his or her ability runs out. Then the person must respond incorrectly to every other item that is more difficult. However, expectations of Guttman model are unrealistically strict as human latent traits are complex and always affected by a multitude of factors (Andrich, 1988). In the Rasch model, unidimensionality is considered to be satisfied when a dominant factor influences the responses to all items. The Rasch model allows uncertainty for responses of persons whose ability is close to item difficulty. This uncertainty results in a non-perfect response pattern around the diagonal of the data matrix as illustrated in Table 2.

Table 1. The response pattern according to the Guttman model

	Items											
Persons	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
a	1	1	1	1	1	1	1	1	1	1	1	0
b	1	1	1	1	1	1	1	1	1	1	0	0
c	1	1	1	1	1	1	1	1	1	0	0	0
d	1	1	1	1	1	1	1	0	0	0	0	0
e	1	1	1	1	1	1	0	0	0	0	0	0
f	1	1	1	1	1	0	0	0	0	0	0	0
g	1	1	1	1	0	0	0	0	0	0	0	0
h	1	1	1	0	0	0	0	0	0	0	0	0
i	1	1	1	0	0	0	0	0	0	0	0	0
j	1	1	0	0	0	0	0	0	0	0	0	0
k	1	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0

Table 2. The response pattern according to the Rasch model

	Items											
Persons	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
a	1	1	1	1	1	1	1	1	1	1	1	0
b	1	1	1	1	0	1	1	1	1	1	0	0
c	1	1	1	1	0	1	1	1	0	0	1	0
d	1	1	1	1	0	1	0	1	1	1	0	0
e	1	1	1	1	0	0	1	0	1	0	1	0
f	1	1	1	1	0	1	1	1	0	0	0	0
g	1	0	1	0	1	1	1	0	1	0	0	0
h	1	1	1	0	1	1	0	0	0	0	0	0
i	0	0	0	0	1	0	0	1	1	1	1	0
j	1	1	0	1	1	0	0	0	0	0	0	0
k	1	0	0	0	1	0	0	0	0	0	0	0
l	0	1	0	0	1	0	0	0	0	0	0	0

The Rasch model is prescriptive in the sense that it formulates the requirements for a fundamental measurement and makes it possible to verify if these requirements are satisfied in the observed data. Whether the data fit the unidimensional Rasch model is an empirical question. The question is, 'Do the data fit the chosen Rasch model?' (Rasch, 1960; Wright and Stone, 1979; Andrich, 1989; Bond and Fox, 2001; Andrich, 2004; McHorney and Monahan, 2004; Conrad and Smith, 2004). From the view of the Rasch model, identifying why item discrimination (the slope of the ICCs) differs and how to control these factors are problems to be solved. The usual practice consists in removing items that discriminate differently (Andrich, 2004; McHorney and Monahan, 2004). Contrary to the Rasch model, the two- and three-parameter logistic models fail to satisfy the property of specific objectivity, challenging their status as measurement models (Fisher, 1994). These models are descriptive rather than prescriptive in the sense that a measurement model is chosen that best fit the data. The question is, 'Does the model fit the data?' (Bond and Fox, 2001; McHorney and Monahan, 2004; Conrad and Smith, 2004).

Rasch analysis offers solutions to practical measurement problems. Among the many applications, Rasch models may be used for linearizing ordinal scores, scale development, item banking, computer adaptive testing, cross-cultural equivalence, optimizing rating scales, etc. Furthermore, Rasch measurement models have been developed for a wide variety of data types. Initially, Georg Rasch proposed his model for dichotomous response formats (not painful/painful) (Rasch, 1960). Later, the principle of the model was expanded to polytomous response formats, that is items graded on more than two response categories (e.g., none, mild, moderate, severe, very severe) (Andrich, 1978b; Andrich, 1978c; Masters, 1982). The model is also applied to situations in which supplementary factors, in addition to item difficulty and subject ability, systematically influence the observed response. For instance, subjects may be rated by different judges or at different times of the day. A notorious example in which performance ratings depend on assessment made by several judges is, for instance, the diving competition at the Olympic Games (Bond and Fox, 2001). Each diver performs a series of dives (the items) and is rated by several judges who probably differ in levels of severity. The aim is to determine the diver with the highest skill, regardless of which dives are executed, and which judges happen to rate them. The 'many-facet' version of the Rasch model, developed by Linacre (1989), successfully models these more complex situations. This model allows for the estimation of differences in severity between judges, and thus eliminates this kind of judge 'bias' from the calibration of items and the measure of examinees. This model enables to place examinee, item and judge estimates on one common scale. The robust nature of Rasch models in the face of missing data does not require that every examinee be rated by every judge on every item. Sufficiently accurate estimates of these three 'facets' can be calculated with substantially incomplete data matrix. It is only necessary that observations designed to create a network through which every parameter can be linked to every other parameter, directly or indirectly, by some connecting observation. Linacre (1997) provides quite detailed requirements for minimal judging plan.

Specific aims of the present work

In 1989, McArthur and coworkers were the first to advocate the application of the Rasch model in pain measurement. However, until recently, its application in pain measurement was rather limited (Thomeé, 1995; Tesio, 1997; Wolfe, 2003; Pesudov and Noble, 2005). The overall aim of the present work is to demonstrate the usefulness and the efficacy of the Rasch model to provide solutions to a variety of measurement problems in the specific context of pain assessment. As Rasch models rest on the requirement of unidimensionality and as verbal self-report has been described as the most valid measure of the experience of pain (Melzack and Katz, 2006), all chapters of the present work focus on the measurement of a single dimension of the pain experience, its intensity and on self-reported rating scales assessing this dimension. Nevertheless, the Rasch methodology can be applied to the measurement of others dimensions of pain (e.g., unpleasantness) and to other types of responses (e.g., behavioural responses).

An exhaustive overview of the measurement framework provided by the Rasch model will not be presented since the family of Rasch models has previously been described in several reference books (Wright and Stone, 1979; Wright and Masters, 1982; Andrich, 1988; Bond and Fox, 2001; Penta et al., 2005). The main principles and statistical indexes are explained in the methods sections of the different chapters and a theoretical section (section 3.1.) will be devoted to the detailed description of indexes specifically related to the functioning of rating scales. Chapters of the present work center on different uses of the Rasch model in the context of pain measurement: the transformation of ordinal scores into interval-level measures, the development and validation of a new measurement tool and the investigation of rating scale functioning. The content of these chapters is summarized in the lines that follow.

Chapter 1 describes one model of the Rasch family of models, the many-facet model, and illustrates how it can be applied to the analysis of responses on a pain intensity verbal rating scale following experimental heat stimuli. This rather methodological chapter focus on issues of ordinal scores obtained on such scales and on benefits of transforming them into interval-level measures. Advantages and disadvantages of the Rasch methodology are discussed. Chapter 1 is presented as it has been accepted for publication in the *European Journal of Pain*.

Chapter 2 presents the development and validation of the Situational Pain Scale (SPS), a measure of the mental representation of pain in imaginary painful situations, in a population of healthy subjects and chronic pain patients. In patients, an examination of the relationship between the SPS measure and other clinical and psychological variables was conducted. The measure of pain representation was significantly related to present pain intensity and anxiety about pain. Fibromyalgia patients showed significantly higher pain representation measures than all other groups. The development and validation of the SPS and the identification of variables that correlate with this instrument are presented as it has been submitted for publication.

Chapter 3 examines the functioning of rating scale categories and more specifically the question of the optimal number of response levels required to assess the intensity of pain. A first section details the problematic, reviews studies previously conducted and presents the statistical indexes provided by the Rasch methodology for investigating the functioning of rating scales. Two empirical studies follow. They investigate the optimal number of response levels by means of the Rasch analysis. The first one is devoted to the particular case of verbal rating scales in healthy adults and the second one to faces scales in healthy children aged from four to seven years. The latter is presented as it has been submitted for publication.

Finally, the last section discusses the results of the different chapters and presents future perspectives.

Chapter 1
Measure of experimental pain using Rasch analysis

Abstract

Most common instruments used to assess the painfulness of nociceptive stimuli and the perception of such stimuli are ordinal. This property limits arithmetic operations and statistical procedures that can be applied on their numbers. The Rasch methodology provides mathematical procedures for transforming scores on an ordinal scale into measures on an interval scale. The present paper aims at presenting the basics of this methodology by applying it to the measurement of experimentally induced pain. Six blocks of seven CO_2 laser heat stimuli varying in intensity were delivered on the hand of one hundred healthy subjects. They rated their pain perception on a three-level verbal rating scale (not painful – slightly painful – painful). One member of the family of Rasch models, the many-facet model, was applied to the analysis of these ratings. The analysis provided linear measures of the painfulness for each intensity of stimulation, of the pain perception of each subject and of the painfulness of each successive block. All these measures are located on a single pain perception continuum. Advantages and disadvantages of this methodology will be discussed in terms of subsequent possible mathematical analyses, statistical tests and implications for experimental and clinical investigations.

Accepted for publication as:

Decruynaere C, Thonnard J-L, Plaghki L. Measure of experimental pain using Rasch analysis. Eur J Pain 2007;11:469-474.

1. Introduction

Most common instruments used to assess the painfulness of nociceptive stimuli and the perception of such stimuli are ordered verbal or numerical rating scales. These scales are widely used as they are inexpensive, easy to design and to administer, and because they provide quite valuable information. Numbers are assigned to the different response categories. Ordinal scores usually start at zero for the lowest category and follow a regular progression by increment of one till the highest category. Nevertheless, there is no real basis for choosing one number progression instead of another. Any scheme can be used to assign numbers, as long as the numbers get larger with consecutive categories. These numbers only indicate an ordering relationship and cannot be considered as measures (Merbitz et al., 1989; Wright and Linacre, 1989) because no information is given concerning the distance that separates each pair of adjacent categories of the underlying attribute that is measured. All that is known is that higher numbers represent 'more' of the attribute. Considering a 4-level rating scale: not painful, slightly painful, moderately painful, extremely painful, scored 0, 1, 2, 3 respectively, a reduction from category 2 to category 0 represent a greater relief than a change from category 2 to category 1 but not necessarily twice as much. Often, however, arithmetic operations and parametrical statistics are performed on scores obtained on these ordinal scales. These operations are in principle not valid because scores allocated to each category are not true numbers. Consequently, results are potentially erroneous (Townsend and Ashby, 1984). These operations require scales with known and equal intervals. Such scales are defined as interval scales according to the classification of measurement scales established by Stevens (1946). In these scales, the difference between two graduations is known and constant all along the dimension of interest. Compared with ordinal scales, besides the ordering relationship, numbers on interval scales tell us how much more of the attribute of interest is present. Interval scales are linear and quantitative. They allow all elementary arithmetic operations and accordingly calculation of common statistical indicators leading to meaningful quantitative comparisons within and between subjects.

In the context of pain measurement, more information could be gained if pain perception was expressed on an interval scale. The family of Rasch models provides mathematical procedures for transforming responses on an ordinal scale into measures of pain perception on an interval scale of measurement. These models take into account differences in stimulus characteristics as well as variations between subjects and allow strong predictions of a subject's future pattern of responses to painful stimuli (McArthur et al., 1992). The present paper aims at presenting the basics of one of these models, the many-facet model (Linacre, 1989), by applying it to the measurement of experimentally induced pain. Finally, advantages and disadvantages as compared to the use of ordinal scores will be discussed in terms of subsequent possible mathematical analyses, statistical tests and implications for clinical investigations.

2. Methods

2.1. Subjects

Experiments were performed on 100 healthy volunteers. These subjects had no history of neurological, psychiatric or chronic pain disorders. This was determined on the basis of a short interview at the time of the recruitment. Eight participants were excluded from the analysis for technical reasons. A significant thermal drift of the laser power occurred during their experimental session. The 92 remaining subjects (39 males and 53 females) were between 20 and 59 years of age (mean: 39.4). The study was authorized by the ethics committee of the Université catholique de Louvain, Faculty of Medicine in Brussels, Belgium.

2.2. Experimental apparatus and design

Laser stimulator

Cutaneous heat stimuli were delivered by a CO_2 laser (10.6 μm wavelength) (Plaghki et al., 1994) on the dorsum of the right hand. The laser was designed and built in the Department of Physics (Université catholique de Louvain, Belgium). Stimulus duration was 50 ms and its surface area was 79 mm² (10 mm diameter). In order to minimize habituation or nociceptor sensitization, the laser beam was moved between each stimulus

Experimental design

A few stimuli varying in strength were applied before starting the recording session in order to familiarize subjects with the nociceptive stimuli. The recording session was divided into 6 successive blocks. Each block consisted of 7 different test stimuli (325, 400, 475, 550, 625, 700 and 775 mJ) applied in random order with an inter-stimulus interval of 10-15 seconds. A session lasted about 15 minutes.

Data acquisition: pain perception

Subjects were informed that they would receive stimuli of varying laser intensities, which could produce tactile and/or thermal sensations like a touch, a tingle, a pinprick, warmth or a burn. They were instructed to rate the intensity of their perception, whatever the quality of the sensation, on a three-level verbal rating scale labeled 'not painful', 'slightly painful', 'painful' after each laser stimulus. The scale appeared on a computer screen placed in front of the subject who had to check one of the three categories with a mouse hold in the left hand. Stimuli that were not perceived were considered as not painful. The response categories (not painful, slightly painful, painful) were scored 0, 1 and 2 respectively so that subject's total scores (TS_{subj}) could range from 0 to 84 (2 * 7 stimuli * 6 blocks). From the 92 subjects, seven always answered in the category 'not painful' providing an extreme TS_{subj} score of 0. As extreme scores imply undefined measures from the Rasch analysis, these 7 subjects were not taken into account for the subsequent analysis. The total score for a given stimulus intensity (TS_{stim}) is the sum of the scores of all subjects through all blocks. TS_{stim} can range from 0 to 1020 (2 * 85 subjects * 6 blocks). The total score for a given block (TS_{block}) is the sum of the scores of all subjects to all stimuli. TS_{block} could range from 0 to 1190 (2 * 85 subjects * 7 intensities).

2.3. Rasch analysis

Developed in the 1950's, the Rasch model is part of a family of models known as item response theory (IRT). This model requires that only stimulus painfulness and subject's perception determine the probabilities of response category choice when subject has to score pain evoked by the stimulus. These two parameters (subject's perception and stimulus painfulness) are estimated by the model from the matrix of reported responses. Initially, the model was developed for dichotomous data (e.g., not painful/painful) (Rasch, 1960). Later, the original model has been extended to the analysis of polytomous items, i.e., with more than two ordered response categories (Wright and Masters, 1982; Andrich, 1988). In the present design, with the inclusion of repetitive blocks, it is useful to estimate simultaneously not only the perception of the subject and the painfulness of the stimulus but also the block painfulness allowing to quantify the effect of block repetition on pain perception. This is accomplished by applying another expansion of Rasch's original (i.e., two-facets) model called the many-facet model (Linacre, 1989).

Responses to the experimental stimuli were thus considered as a three-faceted data matrix (85 subjects * 6 blocks * 7 intensities of stimulation). A many-facet Rasch analysis was performed using FACETS© (Linacre, 1994a). The rating scores were analyzed with the three-faceted rating scale model as formulated below (Linacre, 1989):

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \beta_n - \delta_i - \lambda_j - \tau_k$$

Where

P_{nijk} is the probability of subject n giving a rating of k on stimulus i in bloc j

P_{nijk-1} is the probability of subject n giving a rating of k-1on stimulus i at the bloc j

β_n is the pain perception of subject n (n = 1, 2, …, 85)

δ_i is the painfulness of stimulus i (i = 1, 2, …, 7)

λ_j is the painfulness of block j (j = 1, 2, …, 6)

τ_k is the threshold between each pair of adjacent categories and k = 1, 2.

This formula computes the probability that a subject will give a particular response to a given stimulus in a given block. As the rating scale used in this study is polytomous (three response categories), the model also provides a measure of the thresholds (τ_k) that separate each pair of adjacent response categories. Subject, stimulus, block and threshold measures are accompanied by standard errors (SE) representing the range within which the true measures are expected to lie. All these measures are located on the same equal-interval scale and are expressed in the same unit, the logit. The logit actually expresses a difference (e.g., between subjects, between stimuli or between subjects and stimuli). One logit can be defined as the increase in subject's pain perception that increases the probability of responding in a given category rather than in the category below by a factor of $e^1=2.71$. As stimuli are located on the same measurement scale, one logit can also be defined as the increase in stimulus painfulness that increases the probability of responding in a given category rather than in the category below by a factor of $e^1=2.71$. This unit of measurement is constant all along the pain perception scale.

Rasch analysis estimates separation reliability for subjects, stimuli and blocks (Wright and Masters, 1982; Fisher, 1992). The index of separation reliability (R) is defined as the ratio between the true measure variance and the observed (true + error) measure variance. In other words, it corresponds to the proportion of observed variance that is not due to measurement error. Separation reliability indicates the reliability with which measures can be separated, with values ranging between 0 and 1 (Wright and Masters, 1982). The higher R, the better the measures are separated. High subject separation reliability means that subjects are well separated according to their pain perception; high stimulus separation reliability means that stimuli are well separated according to their painfulness; high block separation reliability means that identical intensities of stimulation are differently perceived from one block to another.

3. Results

The dominant tendency among subjects was to use low pain ratings for laser stimuli as the percentage of responses to all stimuli and all blocks in the different categories was 60%, 30% and 10% for 'not painful', 'slightly painful' and 'painful', respectively. Stimuli that were not perceived (11%) were considered as not painful (cf. methods). As expected, we observed an increase in the TS_{stim} with the increase in the laser power (Table 1, two first columns). Concerning the TS_{block}, we observed a slight decline from the first block to the sixth (Table 2, two first columns) but this decrease was not perfectly monotonic.

The many-facet Rasch analysis that follows allowed us to quantify the pain perception. This analysis resulted in linear measures in logit units for subjects, stimuli, blocks and thresholds. All these measures can be located on a single pain perception continuum.

3.1. Stimulus measure

Stimulus measures are reported in Table 1 in order of increasing painfulness (range: 2.97 to –2.32 logits), with greater logit values indicating less painful stimuli. The scale was centred (zero logit) to the sum of the painfulness of all stimuli. Table 1 also presents the SE of the stimulus measure. Reliability index (R) for stimuli approximates 1.00, meaning that almost 100% of the observed variance resulted from true measure variance.

Table 1. Stimulus measure

Stimulus (mJ)	TS_{stim}	Measure (logits)	SE (logits)
325	24	2.97	0.22
400	65	1.76	0.14
475	129	0.80	0.11
550	232	-0.21	0.09
625	367	-1.21	0.08
700	454	-1.79	0.08
775	539	-2.32	0.08

3.2. Block measure

Measures and SE for the six blocks are presented in Table 2, with greater logit values indicating that blocks are perceived as less painful. Blocks tend to be perceived as less painful over time. However, that decrease in block painfulness was rather small, ranging from -0.21 to 0.24 logit (range = 0.45 logit) and not wholly monotonic (see last block). Moreover, if we consider the SE associated with the estimation of the block measure, we cannot conclude to a significant block effect. Indeed, the SE signifies that you have 95% of chance that the true block measure lies within a range of approximately ± 2 SE. This range is called the confidence interval (CI). The CIs of five blocks out of six overlap making this block effect not significant. Reliability index (R) was 0.73, meaning that 27% of the observed variance resulted from measurement error.

Table 2. Block measure

Block	TS_{block}	Measure (logits)	SE (logits)
1	327	-0.21	0.09
2	321	-0.17	0.09
3	318	-0.13	0.09
4	286	0.13	0.09
5	273	0.24	0.09
6	285	0.14	0.09

3.3. Subject's measure

Distribution of subjects' measures is displayed on the top panel of Figure 1. These measures range approximately from –6 to 1.5 logits with higher values associated with higher pain reports. Complete table with measures and SE is not displayed. Subject separation reliability of 0.93 indicates that the calibrated stimuli have well spread out the subjects along the pain perception continuum.

3.4. Relationship between the subject's measure and its total score (TS_{subj})

The bottom panel of Figure 1 shows the ogival relationship between the TS_{subj} (range: 0-84) and the measures of pain perception. Note that this relationship is almost linear between approximately –2 and 2 logits. Nevertheless, this relationship is not linear outside this range where half of the subjects are located (Figure 1, top panel).

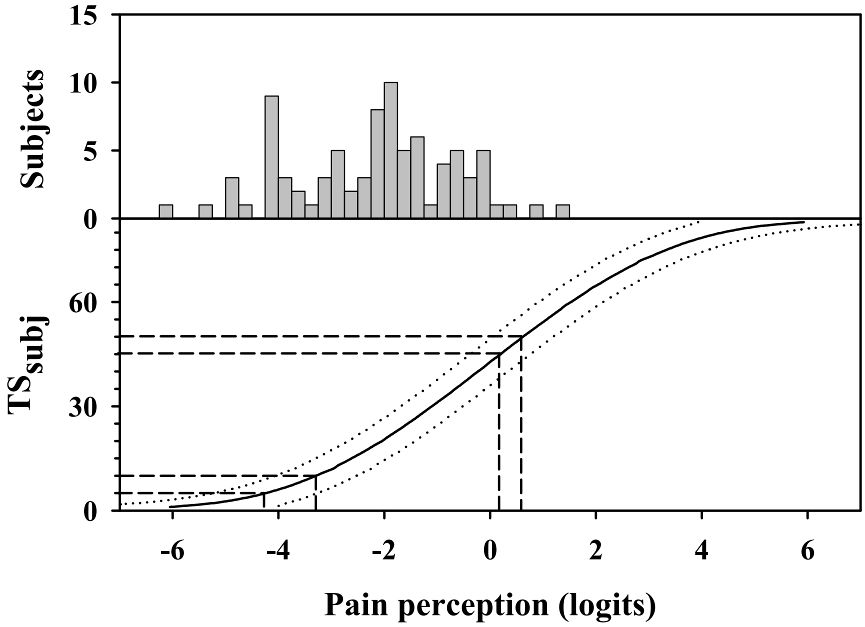


Figure 1. Top panel: distribution of subjects' pain perception measures. Bottom panel: relationship between subject's total scores (TS_{subj}) (range: 0-84) and pain perception measures (solid line) and its 95% confidence interval (dotted lines). The ogival shape of the relationship accounts for the non-linear transformation of scores into linear pain perception measures. The broken lines illustrates that an identical difference in terms of score (5 points) all along the scale does not necessarily represent an identical difference in terms of measure.

3.5. The most probable response and threshold measure

The top panel of Figure 2 is the item map for the first block. This item map displays the most probable response to a given stimulus as a function of the measure of pain perception. By comparing the measure of a given subject to the painfulness of each stimulus, it is possible to determine his most probable response category to each stimulus of this first block. For instance, a subject with a perception of –2 logits (dotted line in Figure 2) would be expected to rate the four less painful stimuli as not painful and the three most painful as slightly painful. None of the stimuli would be rated as painful. On the item map, the points of transition from one grey shade to the adjacent one are the thresholds of the first block. For a given stimulus, the threshold between two consecutive response categories corresponds to the level of perception for which the response probabilities of the two adjacent categories are equal. The item map for the other blocks can be obtained simply by translating to the right the thresholds by the difference of measure between the first block and each subsequent block (0.04, 0.08, 0.34, 0.45 or 0.35 logit, see Table 2, third column). As a fact, the shift is quite small.

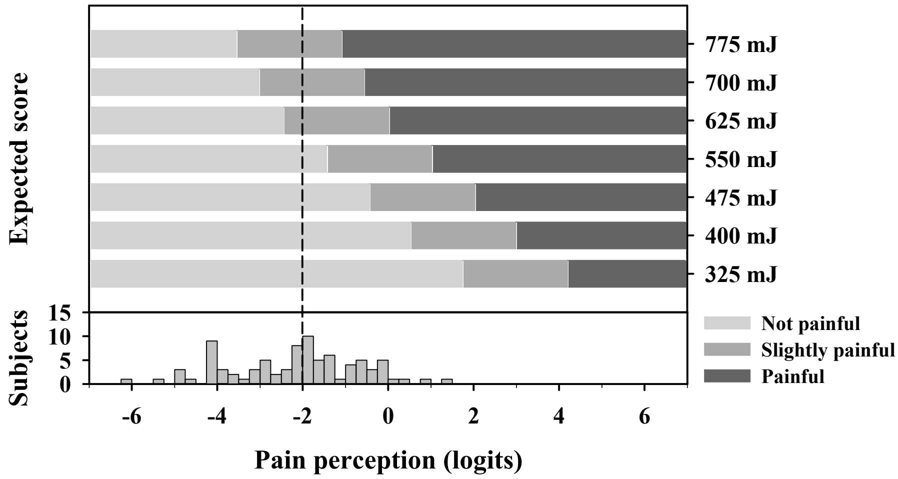


Figure. 2. Top panel: item map for the first block providing a subject's most probable response to each stimulus as a function of the measure of pain perception. Bottom panel: distribution of subjects' pain perception measures.

4. Discussion

The purpose of the present study was to present a probabilistic model, the many-facet Rasch model, to analyze rating scale data and to apply it with experimental pain ratings. This model overcomes the limitations of ordinal scales by transforming scores of the different parameters (subjects, stimuli and blocks in the present case) into measures on a common interval scale. This psychometric approach, derived from item response theory (IRT) is still not much exploited by the pain community perhaps due to its "originality and specific terminology" (Tesio, 2003). Few studies have applied the Rasch model to quantify pain perception in experimental (McArthur et al., 1992) and clinical settings (e.g., McArthur et al., 1991; Thomeé et al., 1995; Tesio et al., 1997; Wolfe, 2003; Pesudovs and Noble, 2005).

A major advantage of Rasch analysis consists in the resulting linear measures of pain perception for subjects. This allows comparing quantitatively either two subjects (or two groups of subjects) or the same subjects at two different moments. A difference of 1 logit unit between two subject's measure indicates that the subject with the higher perception has a probability of responding in a given category (e.g., 'slightly painful') rather than in the category below (e.g., "not painful") raised by a factor e^1=2.71 whatever the stimulus, the block and their absolute location along the scale. The ogival relationship between TS_{subj} and subject's measure estimated by the model (Fig. 1, bottom panel) means that in the central part of the scale, this relationship is approximately linear but outside the middle of the scale, it is not. In other words, according to the location along the scale, an identical increment in the TS_{subj} (on the y-axis) does not represent an identical increment in pain perception measure (on the x-axis) all along the pain perception continuum. The broken lines illustrate that point. Consider two subjects whose TS_{subj} equals 50 and 10 respectively. An equal decrease of 5 points in terms of TS_{subj} for both subjects corresponds to very unequal pain relief in terms of measure of pain perception expressed in logit units (0.42 versus 0.95). The more you move away from the centre of the scale (TS_{subj} of 42), the more a difference of one in terms of TS_{subj} represents a large variation of measure of pain perception. This can be particularly critical as illustrated in this study where half the population is located on the left of the linear part of the relationship. In clinical practice, when dealing with patients we can suppose that subjects may often be located at the extremities of scales where TS_{subj} are not linear. Quantitative comparisons using TS_{subj} for these subjects could lead to erroneous conclusions. In the present case, we may attribute the relatively low pain perception to the reluctance of healthy subjects to label a very brief and well-controlled stimulus as painful particularly in the absence of any anxiogenic component (subjects knew that the experiment did not carry any risk). Higher stimulus intensities were not used due to the potential risk of skin burn.

Concerning the block parameter, the many-faceted Rasch analysis resulted in slightly different measures for the successive blocks with a tendency to be perceived as less painful over time. However, this effect was not significant, providing evidence for the absence of habituation or sensitisation. It is worthwhile to notice here that this many-faceted model can also be used to quantify the effect of experimental or therapeutic interventions. For instance, if we had to test an analgesic drug, this method could have revealed the presence or absence of a treatment effect and, most importantly, quantified it on a linear scale.

Of course, the limitations of the Rasch analysis must also be taken into consideration. The first limitation concerns the intellectual investment in order to fully understand the model as well as to handle the analysis software. Indeed, the analysis of ratings with the Rasch model requires more effort than the traditional handling of ordinal scores. The second limitation is related to the sample size required for building a high precision instrument for measurement. In general, the bigger the sample size, the more precise the stimulus calibration estimated by the model. This measurement precision is modelled by the SE associated with the estimated stimulus measure. In the present study, the stimulus SE was around 0.1 logit meaning that we have 95% confidence that stimulus calibration was no more than 0.2 logit away from their estimated value (Linacre, 1994b). Notice that the SE for the lowest intensities of stimulation was higher (Table 1). This is explained by the poorer targeting of these intensities with the sample of subjects. Moreover, the scale calibration holds only for subjects presenting the same characteristics than those of the sample used to calibrate the scale. It means that for clinical applications, invariance of the measurement instrument must be verified according to clinical state, age and gender.

However, these limitations should not constitute an obstacle for using the Rasch model. They can be overcome. Indeed, once an investigator sufficiently familiarized with the Rasch model has calibrated the measurement scale, the relationship between the subject's total score and its pain perception measure is determined. Consequently, each time another investigator, maybe less familiarized with the model, assesses a single additional subject, he can use the total score to obtain the linear measure of pain perception by simple graphical interpolation of this relationship (Figure 1, bottom panel).

Finally, two advantages of the Rasch model must be underlined. First, it allows obtaining measures of perception from a relatively small number of rating data per subject. Although, long experimental designs are not a major concern in the experimental field, they can become one in the clinical domain where timeconsuming procedures are not realistically done neither for the patient nor for the practitioner. In the present study, in view of the absence of a meaningful block effect, we can envisage to further reduce the number of blocks. The only disadvantage of this reduction lies in the increase of the SE of subjects' measure or in other words in the decrease in subjects' measurement precision. A second advantage of the model is that its application is not limited to experimental stimuli for which we have a known physical calibration. It can also be used in a clinical context when rating data correspond for instance to pain experienced during different conditions such as walking, standing or sitting (McArthur et al., 1991; Tesio et al., 1997; Thomeé et al., 1995; White et al., 2002; Davis et al., 2003).

In conclusion, the Rasch methodology overcomes the limitations associated with the use of ordinal scores. It provides scientific rigorousness to pain perception assessment, which can then be considered as linear, quantitative information rather than "soft" qualitative data (Tesio, 2004). Therefore, this method contributes to eliminate the difference existing between clinical variables such as pain and physical or biological variables in terms of measurement. Finally, contrary to ordinal pain scores, Rasch measures can validly be treated with parametric statistical procedures with all the subsequent advantages in terms of power.

Chapter 2
The Situational Pain Scale: a measure of pain representation in imaginary painful situations

Abstract

In the present study, an instrument, the Situational Pain Scale (SPS), was devised to measure the mental representation of pain in imaginary painful situations and validated in a population of healthy subjects and chronic pain patients. A list depicting 43 painful situations was given to 100 healthy subjects and 111 chronic pain patients who rated the pain intensity for each situation on a four-level verbal rating scale. Chronic pain patients also completed questionnaires assessing clinical and emotional attributes. Responses to the SPS were analyzed according to the rating scale Rasch model to select unidimensional and unbiased items. The final SPS consisted of 18 items with a widespread measurement range, a distribution well-targeted for the sample of healthy subjects and chronic pain patients and a high index of measurement precision. The measure of pain representation was significantly related to present pain intensity and anxiety about pain. Fibromyalgia patients showed significantly higher pain representation measures than all other groups. In conclusion, the SPS provides a valid measure of pain representation in imaginary situations.

Submitted for publication as:

Decruynaere C, Thonnard J-L, Plaghki L. The Situational Pain Scale: a measure of pain representation in imaginary painful situations.

1. Introduction

We all regularly experience painful situations. Some situations are very painful (e.g., spilling boiling water on the hands), while others are much less painful (e.g., a mosquito bit). As situations differ in noxious character, perception of pain also differs with each person. For an identical painful situation, some individuals hardly manifest pain whereas others exhibit high levels of pain. The report of this pain perception includes both a sensory-discriminative component reflecting the functioning of the sensory system and an affective-motivational component reflecting the distress and drive generated by a given painful situation. The emotional component is more complex because it comprises several factors, which may be related to one another. For instance, these factors include current emotional status, painful past experiences, socio-cultural contexts, expectations of the interviewer and the social desirability of the answer. The measure of the global effect of these factors, also called the attitude of subjects (Clark, 1974), on pain perception could constitute a useful prognostic indicator of pain behaviors to guide care-givers in their management of both acute and chronic pain patients. For instance, this measure could be used as a predictive tool for detecting which acute pain patients are more at risk of chronification or which preoperative patients will require more attention, care and possibly analgesia in the post-operative period.

The current study describes the development and validation of an instrument, which aims at providing a global measure of the influence of these factors on pain perception. This instrument, the Situational Pain Scale (SPS) portrays everyday life painful situations (e.g., I burn my tongue tasting scorching hot food) that subjects have to rate on a verbal pain rating scale. The effect of psychological factors is thus measured through the mental representation of pain intensity in imaginary painful situations. Since this effect cannot be observed directly it has to be inferred from observed consistency in responses. This procedure was chosen for two reasons. First, given that the painful situations must be imagined and not really experienced by the subjects, the emotional aspect should play an important role in the rating process as compared to the sensory aspect that should be minimized. Second, as subjects are not entirely conscious of the influence of psychological factors, the procedure utilizes an indirect method of assessing the global effect of pain affect via the rating of painful situations. The influence of psychological factors is inferred from the consistency to favour high or low pain scores. This procedure does not require that subjects are aware of factors influencing their pain affect.

Most pain-related assessment scales are developed and validated using the classical test theory, despite the documented shortcomings (Hambleton et al., 1991). One first and major limitation is the ordinal property of scales produced by classical test theory. Scores provided by these scales are not necessarily proportional to the quantitative level of the latent variable. For instance, if we consider a four-level verbal scale, not painful (0), slightly painful (1), moderately painful (2), extremely painful (3), the category labeled 'moderately painful' represents a higher level of pain than the category 'slightly painful' but does not necessarily represent a double level of pain. A second major limitation is the dependency between the subject's pain score and the particular items selected to assess his level of pain. If a questionnaire comprises very painful items such as 'spilling boiling water on the hands', the subject's pain score will be higher than if the questionnaire is comprised by less painful items such as 'a mosquito bite'. In other words, the subject's location along the pain continuum will depend on the painfulness of the selected items. Similarly, a score for the item, 'standing for 1 hour,' will be higher if the sample who answers it is made up of low back pain patients than if it is made up of healthy subjects so the item location along the pain continuum will depend on the particular sample answering it.

For these reasons, alternative test theories and models that overcome these limitations have been developed (Hambleton et al., 1991). This family of logistic models is known as Item Response Theory (IRT). These probabilistic models differ in the number of parameters describing the items. The model selected in the present study is the one-parameter rating scale model (Andrich, 1978b; Wright and Masters, 1982) from the Rasch family of one-parameter models (Rasch, 1960). This model prescribes that the probability of endorsing any response category to an item solely depends on the subject's measure of pain representation, the painfulness of the item, and the location of thresholds. Thresholds are the boundaries between contiguous response categories. More formally, they correspond to the location along the pain representation variable for which two adjacent response categories have an equal probability of occurrence. In case of the four-level verbal scale used in the present study, three thresholds are estimated. The measure of pain representation of each subject, the painfulness of each item (item calibration) and the location of thresholds are not known *a priori*. They are estimated from the matrix of reported responses (Wright and Panchapakesan, 1969; Wright and Masters, 1982) and are located on the same measurement scale and expressed in the same unit. The basic requirement of Rasch models is that the set of items provides a measure of one single attribute (or dimension) at a time, unbiased by other qualities of the subject being measured. This requirement is called the requirement of unidimensionality and is considered to be satisfied when a dominant attribute influences the responses to all items (Hambleton et al., 1991).

These models possess the advantage of overcoming the aforementioned limitations of ordinal scales. They provide measures for items, subjects and thresholds on an interval scale. On interval scales, the difference between two graduations is known and constant along the dimension of interest allowing for elementary arithmetic operations. Furthermore, the measure of each subject is estimated independently of the relative painfulness of the particular item set used to collect the responses and, similarly, the painfulness of each item is estimated independently of the level of pain of the particular sample answering the items (Rasch, 1960).

The primary objective of the present study is to build with the rating scale model an instrument to measure the representation of imaginary painful situations. This Situational Pain Scale (SPS) will be validated concurrently in a population of healthy subjects and chronic pain patients. The secondary objective is to identify clinical and psychological variables that covariate with this instrument. Variables assessed concomitantly with the SPS are present pain intensity, current emotional status, trait anxiety and anxiety about pain. The presence or absence of covariation between SPS and these variables will be used for specifying the factors that are associated with the representation of painful situations.

2. Methods

2.1. Subjects

The Situational Pain Scale (SPS) was submitted to a sample of healthy subjects (n = 100) and to a sample of chronic pain patients (n = 111). Healthy subjects (45 males and 55 females), aged 20-59 years (mean: 39, SD: 11.2) had no history of neurological, psychiatric or chronic pain disorder as determined on the basis of a short interview. They were recruited by personal contact of the investigators with people linked (friends and relatives) to the Unité de Réadaptation et de Médecine Physique of the Université catholique de Louvain, Belgium. Chronic pain patients (38 males and 73 females) were between the ages of 26 and 82 years old (mean: 52, SD: 11.6). All patients were recruited in the Multidisciplinary Pain Centre of the Cliniques universitaires Saint Luc in Brussels, Belgium. Patients were referred by general practitioners or specialists. They were examined by an algologist in the Pain Centre. The inclusion criterion was a state of chronic pain for at least a 6-month duration. Exclusion criteria were: (a) current alcohol dependency or abuse; (b) severe psychiatric illness; (c) malignancy; (d) a concurrent acute physical illness (e.g., fractures). On the basis of the examination of the patient and available clinical and laboratory data and according to the International Association for the Study of Pain and classification of chronic pain syndromes (Merskey and Bogduk, 1994), this group was composed of 40 patients with chronic low back and/or radicular pain, 35 with fibromyalgia syndrome, and 36 with other miscellaneous chronic pain conditions (e.g., cervical pain, dorsal pain, neuropathic pain, complex regional pain syndrome or visceral pain syndrome). The duration since the onset of the pain symptoms ranged between 6 months and 50 years.

2.2. Situational Pain Scale (SPS)

2.2.1. Preliminary questionnaire development

Items were assembled to cover the largest range of painful situations. The preliminary questionnaire included 51 items depicting imaginary painful situations; 30 items came from the Situational Pain Questionnaire (Yang et al., 1983) and 21 additional items were devised to extend the range of painfulness. This preliminary version was submitted to a jury of 25 healthy subjects, aged 19-53 years, without particular knowledge of the pain domain, recruited in the same way as the healthy group described in section 2.1. They were asked to estimate the pain they would experience in each situation on a four-level scale (not painful, slightly painful, moderately painful, and extremely painful). Responses were analyzed with a rating scale model (more details are provided in section 2.4.). The first objective of this analysis was to check whether the four-level verbal rating scale was used correctly according to the criteria edited by Linacre (2002), namely the category frequency, the ordering of response categories and the quality of the scale from measurement perspectives (e.g., reliability index). Examination of these criteria revealed that the four-level verbal rating scale was properly used by the sample. The second objective of this preliminary analysis was to perform a first selection on items by removing those that did not contribute to the measurement of the common unidimensional variable of pain representation as determined by the mean-square fit statistic values (more details about the fit analysis are provided in section 2.4.).

2.2.2. Procedures

The experimental version of the SPS included 43 items and was administered as a self-reported questionnaire. A computer generated a pseudo-random sequence of the 43 items and a hard copy was produced for each participant. For each item, subjects were asked to estimate the pain intensity on a 4-level rating scale: not painful (0), slightly painful (1), moderately painful (2) or extremely painful (3). Moreover, subjects had the opportunity to rate situations as ‘impossible to estimate’. This response was encoded as missing data. Test-retest reliability was investigated by resubmitting the questionnaire a second time after a delay of 148 ± 94 days.

2.3. Clinical and emotional assessments

Patients completed a battery of questionnaires designed to assess the following constructs: present pain intensity, current emotional status, trait anxiety and anxiety about pain. These instruments have been validated in chronic pain patients.

Present Pain Intensity. Present Pain Intensity (PPI) was assessed with a visual analogue scale (VAS) consisting of a 150 mm horizontal line on a paper sheet with the words 'no pain' and 'the worst imaginable pain' as the endpoints. Patients were asked to put a mark on the line at the point that best represented their pain intensity at the moment of the evaluation. Its validity and reliability for the assessment of pain intensity in chronic pain patients has been demonstrated (Price et al., 1983). Factor analyses performed on 6 pain intensity measures including the VAS revealed that in a sample of chronic pain patients, the VAS correlated substantially with the first factor, which emerged from the factor analytic procedures; correlations ranged between 0.70 and 0.89 according to the aspect of pain (current pain, the most, the least or the average pain experienced during the past week) (Jensen et al., 1986). Finally, numerous studies conducted in chronic pain patients have shown its sensitivity to treatment effect (e.g., Harke et al., 2005; Zambito et al., 2006).

Current emotional status. Current emotional status was assessed by the Hospital Anxiety and Depression scale (HAD) (Zigmond and Snaith, 1983). This instrument was chosen because it has been explicitly designed for use in medical patients (such as chronic pain patients) as somatic items referring to symptoms of anxiety and depression relating also to physical disorder, such as dizziness, headaches, insomnia, anergia and fatigue, have been excluded from the scale. It comprises two subscales, one assessing anxiety (HAD-A) and one assessing depression (HAD-D). Both subscales contain seven intermingled items assessing patient's feelings during the preceding week. Literature reviews addressing the evaluation of psychometric properties of the HAD support its two-dimensional scale structure, its internal consistency, its concurrent validity, its sensitivity for mild symptoms and change over time and its good acceptance (Herrmann, 1997; Bjelland et al., 2002). Its validity has been demonstrated in chronic low back pain

(Greenough and Fraser, 1991) and it has been repeatedly used in studies involving chronic pain patients (e.g., Ledingham et al., 1993; Gagliese and Melzack, 2003).

Trait anxiety. Trait anxiety was assessed by the trait part of the State-Trait Anxiety Inventory (STAI-T) (Spielberger et al., 1970). Trait anxiety refers to relatively stable individual differences in anxiety proneness, that is, differences in the disposition to perceive a wide range of stimulus situations as dangerous or threatening and the tendency to respond to such threats with anxiety reactions. This scale has extensively been used in studies involving chronic pain patients (e.g., Sedlak, 1985; Gaskin et al., 1992; White et al., 2002).

Anxiety about pain. Anxiety about pain was assessed by the Pain Anxiety Symptoms Scale (PASS) (McCracken et al., 1992). This scale, specifically designed for chronic pain patients, consists of four 10-item subscales: (1) cognitive anxiety (PASS-CA), cognitive symptoms related to the experience of pain such as inability to concentrate, (2) escape-avoidance (PASS-EA), escape and avoidance behaviors related to reducing pain, (3) fearful appraisal of pain (PASS-FA), fearful thoughts related to the experience of pain or negative consequences of pain, (4) physiological anxiety (PASS-PA), physiological symptoms of pain-related anxiety. McCracken et al. (1992) showed that the scale has adequate internal consistency. Cronbach's coefficient alpha ranged between 0.81 and 0.89 for the different subscales and equals 0.94 for the total score. The authors demonstrate a considerable degree of concurrent validity in relation to measures of disability, depression, and medication use, all pervasive consequences of chronic pain.

2.4. Data analysis

2.4.1. The rating scale model

The responses to the SPS were analyzed with the Winsteps Rasch analysis software (Linacre and Wright, 1998) according to a rating scale model (Andrich, 1978b; Wright and Masters, 1982). From the matrix of reported responses and using a joint maximum likelihood estimation procedure, the rating scale model provides the measure of pain representation for each subject, the calibration of the painfulness of each item and the relative location of the three thresholds (Wright and Panchapakesan, 1969; Wright and Masters, 1982). Subject, item and threshold measures are accompanied by standard errors (SE), which describe the range within which the actual subject, item and threshold measures are expected to lie.

2.4.2. Item selection

Beginning with 43 items, indices reported from successive rating scale analyses were used to select the items that constituted the final SPS. Items were removed one-by-one, as the suppression of one item makes the scale calibration change. Criteria used to select items were (a) the fit of the data to the unidimensional rating scale model and (b) the invariance of the items according to demographic and clinical subgroups. An item that did not meet one of these two criteria was eliminated.

a) All items must fit a unidimensional scale

Rasch models rest on the basic assumption of unidimensionality and make it possible to verify whether the requirement is satisfied in the observed data. The Rasch model implies a predicted scoring pattern for each subject on all items computed by the model on the basis of the estimated measures of subjects, items and thresholds. The similarity between the observed scores and the predicted responses is reported by the software through two fit statistics (Wright and Masters, 1982; Smith, 2000): (1) the outlier-sensitive fit statistic (outfit) and (2) the information-weighted fit statistic (infit). The outfit is more sensitive to unexpected responses from subjects with a measure far from the item measure. The infit is more sensitive to unexpected responses from subjects with a measure close to the item measure. These statistics enable the researcher to determine how closely the items define the common variable that is measured and detect items that did not define the same underlying variable as the others.

Fit statistics (infit and outfit) were used to detect items that did not satisfy the model requirement of unidimensionality. Both fit statistics are expressed under the form of mean squares and have an expected value of 1, with typical acceptable values between 0.7 and 1.3 (Smith, Schumacker, & Bush, 1998). All items presenting fit values outside this range were removed.

b) All item must be invariant among the different demographic and clinical subject subgroups

Once the items satisfied the property of unidimensionality through appropriate fit indices, the invariance of the scale among the different demographic and clinical subgroups was examined (e.g., men *vs.* women). The Rasch models require that the probability of the outcome of the interaction of any subject and item be solely determined by the pain representation of the subject and painfulness of the item. No sample characteristic may systematically influence responses to any item. The potential lack of scale invariance is commonly known as "item bias" and more recently as "differential item functioning" (DIF) (Smith, 1992). A DIF may induce a systematic misfit to a common scale calibrated for all subjects, and hence constitute a threat to the invariant use of the same measurement scale for all subjects.

The entire sample of subjects was divided into different subgroups on the basis of the following four criteria: (1) group (healthy subjects *vs*. chronic pain patients); (2) sex (male *vs*. female); (3) age (< 47 *vs*. ≥ 47); and (4) situational pain representation (lower *vs*. higher representation). The sample of chronic pain patients was divided into different subgroups on the basis of the following eleven criteria: (5) pathology (low back pain *vs*. fibromyalgia *vs*. other miscellaneous chronic pain conditions); (6) duration since onset of symptoms (≤ 9.2 years *vs*. > 9.2 years); (7) Present Pain Intensity (≤ 49 *vs*. > 49); (8) level of depression scored on the HAD-D scale (< 9 vs. ≥ 9); (9) level of anxiety scored on the HAD-A scale (< 12 *vs*. ≥12); (10) trait anxiety scored on the STAI-T (< 51 *vs*. ≥ 51); (11) anxiety about pain PASS total score (< 95 *vs*. ≥ 95); (12) cognitive anxiety PASS subscore (PASS-CA) (≤ 33 *vs*. > 33); (13) escape avoidance PASS subscore (PASS-EA) (≤ 23 *vs*. > 23); (14) fearful appraisal PASS subscore (PASS-FA) (< 20 *vs*. ≥ 20); (15) physiological anxiety PASS subscore (PASS-PA) (≤ 19 *vs*. > 19). Except for the criteria of group, sex and pathology, a median split was used to separate the sample. Separate analyses were performed on the different subgroups and a *t* test was used to detect any difference in situation painfulness (calibration) among subgroups (Wright and Stone, 1979). This method requires verification that data for each subgroup fits the prescriptions of the Rasch model (Smith, 1992). Items presenting DIF were removed.

2.4.3. Separation reliability

An index of reliability of the overall scale is provided by the Rasch analysis through a separation reliability coefficient (Wright and Masters, 1982; Fisher, 1992; Wright, 1996). The separation reliability coefficient (R) is determined as the ratio between true measure variance and the observed (true + error) measure variance. It corresponds to the proportion of observed variance that is not due to measurement error but to real variation in measure. In other words, this index indicates the reliability with which measures can be separated, with values ranging between 0 and 1 (Wright and Masters, 1982). High (near 1.0) separation reliability is preferred.

2.4.4. Test-retest reliability

Test-retest stability was examined as a further estimate of reliability. From a sample of 211 healthy subjects and chronic pain patients tested at the time of the first assessment, 156 were reassessed (92 healthy subjects and 64 patients). For the healthy subjects, reasons for not reassessing were rooted in the inability to come and complete the questionnaire under the same conditions as those of the first assessment. For the patients, the reason was the absence of visit in the Multidisciplinary Pain Centre within the given time of the study. Test-retest stability for subjects' representation was determined by the intraclass correlation coefficient. The invariance of the scale across the first and second assessment was also tested through a DIF test of both item calibrations.

2.4.5. Relationship between the SPS measures and other clinical indicators

The strength of association between SPS and demographic, clinical and emotional variables was tested with either comparison of means tests (for nominal variables) or correlation coefficients (for continuous variables). A significance level of 0.01 was used. Finally, potential differences in subject's pain representation according to the different groups were examined (healthy, low back pain, fibromyalgia and miscellaneous chronic pain conditions) to study the sensitivity of the scale to differentiate underlying painful pathologies.

3. Results

3.1. SPS validation

3.1.1. SPS calibration

The first criterion for the item selection was unidimensionality. The Rasch analysis of the 43 original situations showed that 14 items presented either infit or outfit mean-squares higher than 1.3 or lower than 0.7. This suggests that these items did not contribute to define the same unidimensional scale as the other 29. The second criterion for the item selection was the invariance of items among the different subject subgroups. The Rash analysis of the remaining 29 items showed that 11 situations presented a DIF. These 11 items were deleted.

Thus, the Rasch analysis of subjects' responses resulted in an 18-item questionnaire. Items are presented in Table 1 in the order of situation painfulness. The scale is calibrated in logits, the natural logarithm of the odds for perceiving a particular situation as painful. The odds represent the ratio of the probability of rating the situation in one category to the probability of rating the same situation in the category just below (e.g., moderately painful *vs*. slightly painful). Consequently, subjects who report high pain intensity ratings have measures of greater magnitude and are located on the right of the scale (see below).

Table 1. The SPS calibration

Item	Measure	SE	Mean square fit statistics	
	(Logits)	(Logits)	infit	outfit
a. My lips are chapped	1.96	0.12	1.12	1.06
b. I get shampoo in my eye	1.45	0.12	1.00	0.95
c. I get a speck of dust in the eye	1.38	0.12	1.20	1.19
d. Someone pulls my hair	0.92	0.12	1.00	0.98
e. I catch my finger in a zipper	0.87	0.12	1.06	1.03
f. I cut myself with a sheet of paper	0.81	0.12	1.03	1.01
g. I get sunburned on my face	0.81	0.11	1.03	1.02
h. I have a splinter under the skin of one finger	0.77	0.11	1.12	1.14
i. I disinfect a sore	0.60	0.11	1.10	1.12
j. I get an injection in the arm	-0.06	0.11	1.12	1.11
k. I walk on burning sand	-0.27	0.11	1.16	1.17
l. I get sunburned and someone touches me on that spot	-0.50	0.11	0.96	0.99
m. I burn my tongue tasting scorching hot food	-0.51	0.11	0.76	0.77
n. I bite my tongue	-0.58	0.11	0.77	0.79
o. I hit my funny bone	-0.71	0.11	1.04	1.03
p. I knock my head on the corner of a piece of furniture	-1.48	0.11	0.75	0.77
q. I stub my toe on a chair leg	-1.78	0.11	0.91	0.93
r. I get my fingers caught in the car door	-3.67	0.16	0.86	0.90
Mean	0.00	0.12	1.00	1.00
S.D.	1.33	0.01	0.14	0.13

3.1.2. Description of the SPS

The SPS is depicted in Figure 1. The two top panels show the distribution of pain representation measures for healthy subjects and chronic pain patients. These measures range from approximately –3 to 4 logits, with higher values associated with higher pain intensity reports. The measures of pain representation are obtained by converting the ordinal total scores on the 18 items into linear measures. The bottom panel illustrates the ogival relationship between the total raw scores (range: 0-54) and the measures of situational pain representation. The middle panel displays the expected response to a given item as a function of the measure of pain representation. By comparing the measure of a given subject to the painfulness of each item, it is possible to determine the most probable response of the subject to each item. For instance, a subject with a pain representation of -2 logits would be expected to rate the nine less painful situations as not painful, the most painful as moderately painful and the others as slightly painful. None of the situations would be rated as extremely painful.

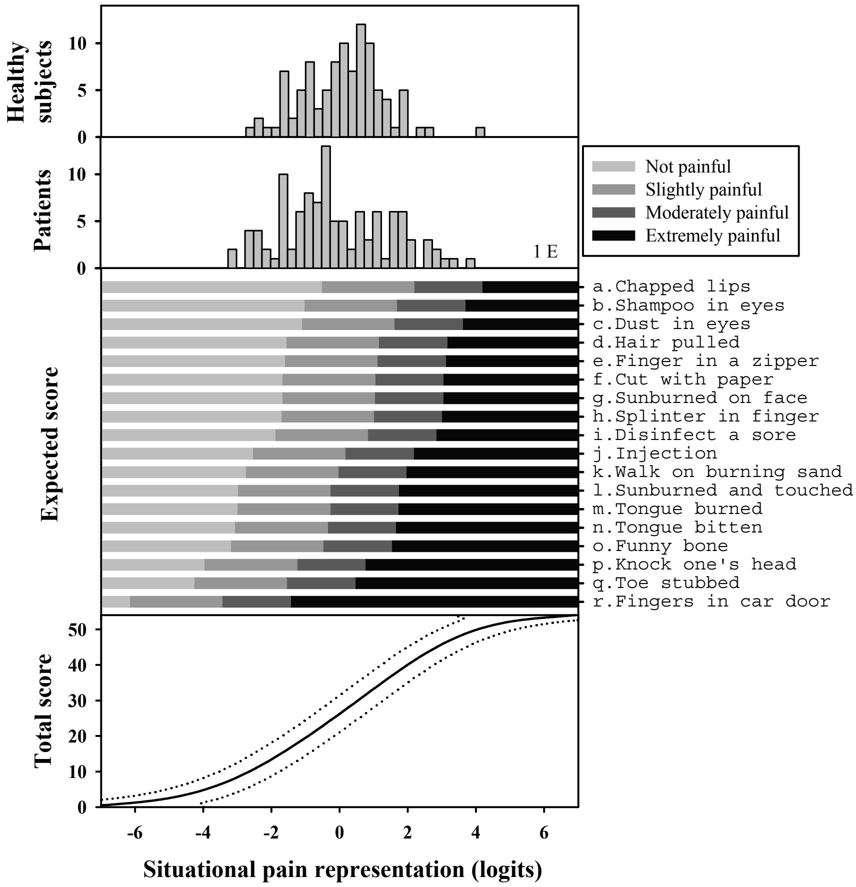


Figure 1. Two top panels: distribution of the SPS measures of healthy subjects and chronic pain patients. One fibromyalgia patient could not be measured by the SPS because all situations were rated as extremely painful (1 E). Middle panel: the item map providing a subject's expected score to each item as a function of the measure of his pain representation. Bottom panel: the relationship between raw scores and measures (solid line) and its 95% confidence interval (dotted lines).

3.1.3. Metric properties of the SPS

The pain intensity measure for the 18 retained items is presented in Table 1 in the order of increasing painfulness (range: 1.96 to –3.67 logits), with higher logit values indicating less painful situations. Standard errors (SE) on these calibrations ranged from 0.11 to 0.16 logit. The mean-square fit statistics indicated that all 18 items contribute significantly to the measurement of pain representation of imaginary situations. The overall scale precision is summarized by a good person separation reliability of 0.92, indicating a good separation of subjects according to their pain representation. The test-retest stability (delay: 148 ± 94 days) of subject measures is presented in Figure 2 (right panel). Subjects' measures estimated by the model at the first and the second assessment are significantly correlated (intraclass correlation coefficient = 0.83, $p < 0.001$). Most of the measures lie within the 95% confidence interval of the identity line indicating that subjects tend to consistently rate the pain intensity of the different situations from one assessment to the other and hence, adopt a quite stable representation over time. The test-retest stability of item measures is presented in Figure 2 (left panel). The painfulness hierarchy of the 18 items is maintained between the first and the second assessment indicating that the scale is invariant over time (intraclass correlation coefficient = 0.99, $p < 0.001$).

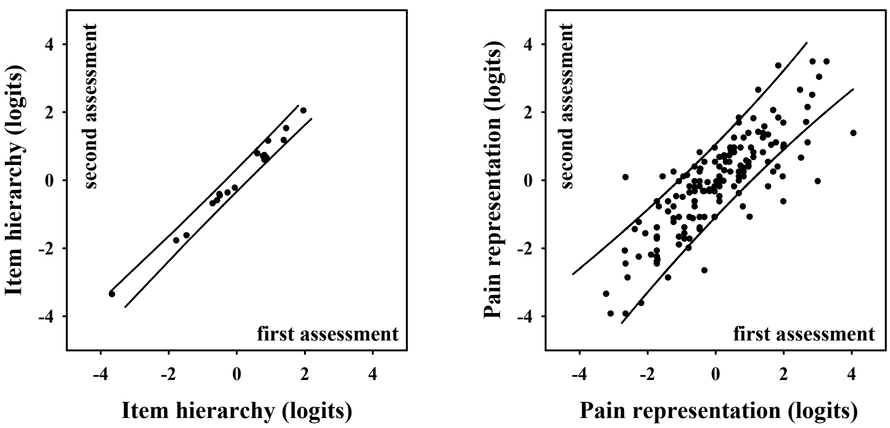


Figure 2. Left panel: differential item functioning plot of the item pain intensity across time and the 95% confidence interval (solid lines) of the ideal invariance. Most painful items are plotted in the bottom/left part of the panel. All items (dots) are lying within the 95% confidence interval indicating that they have identical estimated painfulness across time. Right panel: relationship between the pain representation measures of subject at the first and the second assessment (delay: 148 ± 94 days) and the 95% confidence interval (solid lines) of the ideal invariance. Subjects with a high representation of pain intensity are plotted in the top/right part of the panel. Subject's measures (dots) lying within the 95% confidence interval lines have the same estimated pain representation at the first and the second assessment.

3.2. Relationship between the SPS measures and other clinical indicators

The strength of association between demographic, clinical and emotional variables with SPS measures is displayed in Table 2. No significant differences in SPS measures were observed across demographic indicators (sex and age) or between healthy subjects and chronic pain patients. The duration since the onset of symptoms, current anxiety and depression levels measured by the HAD scale and trait anxiety measured by the STAI-T were not significantly associated with SPS measures. Present Pain Intensity and scores on the PASS, both total score and subscale scores, were significantly related to SPS measures.

Table 2. Relationship between SPS Measures and other Indicators

Variable	Statistic*	*P*
Sex	t = -0.541	0.589
Age	R = -0.025	0.722
Group (patients *vs.* healthy)	t = -0.262	0.794
Pathology (LBP *vs.* FM *vs.* misc. *vs.* healthy)	H = 22.583, df = 3	< 0.001
Duration since onset of symptoms	ρ = -0.068	0.479
PPI	ρ = 0.341	< 0.001
HAD-A	ρ = 0.181	0.058
HAD-D	ρ = 0.150	0.116
STAI-T	ρ = 0.182	0.057
PASS		
Total score	ρ = 0.345	< 0.001
PASS-CA	ρ = 0.280	0.003
PASS-EA	ρ = 0.344	< 0.001
PASS-FA	ρ = 0.244	0.010
PASS-PA	ρ = 0.332	< 0.001

*Reported statistics are as follows: t for t test, H for Kruskal-Wallis test, R for Pearson correlation, ρ for Spearman correlations.
LBP = low back pain; FM = fibromyalgia; misc. = other miscellaneous chronic pain conditions.

Finally, we examined the subject's pain representation according to the different diagnostic groups (healthy, low back pain, fibromyalgia and miscellaneous chronic pain conditions). A significant difference was observed among the groups (ANOVA, $p < 0.001$). Tukey's post-hoc tests revealed that fibromyalgia patients had significantly higher SPS measures than all others groups: fibromyalgia (mean = 0.94, SD = 1.73) *vs*. low back pain patients (mean = -0.57, SD = 1.26, $p < 0.001$), *vs*. miscellaneous other chronic pain patients (mean = -0.49, SD = 1.32, $p < 0.001$) and *vs*. healthy subjects (mean = 0.09, SD = 1.20, $p = 0.006$).

4. Discussion

4.1. SPS Rasch validation

The main purpose of the present study was to develop a measure of pain representation of imaginary painful situations and validate it simultaneously in a population of healthy adults and chronic pain patients. The validation in the healthy population was necessary as it would have been difficult and hazardous to discuss the results observed in chronic pain patients without knowing how healthy subjects perceive imaginary painful situations.

An important preliminary aspect to the study is that the portrayed situations may, besides intensity, evoke other qualities. This brings the unidimensionality of the Situational Pain Scale into question. Pain can be either short (item f) or long-lasting (item m). It can elicit a thermal (item k), mechanical (item q) or pinpricking (item j) sensation and can be localized (item j) or extended (item l). However, statistical indices (infit and outfit statistics) provided by the Rasch analysis confirmed that the 18 items retained for the final version of the SPS measure a unidimensional attribute. Furthermore, the concept of unidimensionality, as considered by Andrich (1988), is a relative matter. For instance, Tesio and coworkers (1997) linked pain and disability into a unidimensional pain/disability variable called "back illness".

The 18-item list shows a continuous progression in painfulness and is well-targeted to the sample of subjects. The scale provides a precise measure, as reported by the person separation reliability of 0.92 in this sample. The invariance of the scale across healthy subjects and chronic pain patients allows its use in both populations. The reproducibility in the item hierarchy between both assessments indicates that the SPS is invariant over time.

The Rasch methodology has been selected for validating the SPS because it is particularly appropriate in providing a measure of the underlying continuum of pain representation. McArthur, Cohen, and Schandler (1989) highlighted that the Rasch model fulfils the criteria of an ideal measurement system as stated by Gracely and Dubner (1981). Based on the responses of subjects in the sample to the items of the scale, the probabilistic Rasch model builds a pain continuum like a measurement ruler whose graduations are defined by the painfulness of the situations. The subject's measure is provided through his/her responses to the set of calibrated situations. A Rasch-built scale has the major advantage of providing an interval-level measurement that overcomes the limitations related to the use of ordinal raw scores. The ogival shape of the relationship between the total raw scores and the measures of situational pain representation (Figure 1, bottom panel) highlights that in the central part of the scale, this relationship is approximately linear, but is not outside the middle of the scale. That means that an identical increment in the ordinal total score of the subject does not represent an identical increment in pain representation measure all along the pain continuum. This transformation is particularly relevant for subjects lying at the extremes of the scale. Thus, quantitative comparisons using the subject's total score are not valid. The Rasch transformation of ordinal scores into interval-level measures enables quantitative comparisons from one person or group to another and over time. It allows the use of parametric statistical techniques while only less-powerful, non-parametrical statistics are valid with ordinal data such as conventional pain scores. Moreover, Rasch computer programs supply valuable statistical indices expressing the psychometric qualities of the scale. Most important are the fit indices and the person separation reliability index. The first is to assess the unidimensionality of the scale, an essential property of the measurement scale, to ensure that the measure is unbiased by other variables. The second allows estimating the ability of the scale to distinguish subjects according to the variable of interest. Thus, the Rasch model provides a strong scientific basis for the development and improvement of pain-related scales.

4.2. Relationship between the SPS measure and other clinical indicators

The second objective of the study concerned the identification of clinical and psychological variables that correlate with SPS. The first major observation concerns the almost perfect similarity of both the healthy subjects' and chronic pain patients' distributions. Nevertheless, very large differences exist within both populations from one subject to another, with some consistently reporting all situations as 'not painful' or 'slightly painful' and others reported almost all of them as 'moderately' or 'extremely painful'. Moreover, the stability of the subjects' measure over time indicates that chronic pain patients as well as healthy subjects maintain their rating, which underlies the enduring character of the pain representation. However, within the patient group, there is a dichotomy between fibromyalgia patients who tend to give high scores and low back pain and miscellaneous other chronic pain patients who tend to give lower scores. The combination of these pathological groups results in a distribution of chronic pain patients almost comparable to that of healthy subjects.

Fibromyalgia patients have significantly higher SPS measures than all other groups. A first explanation for these higher measures is the hypervigilance hypothesis (Chapman, 1986; Gibson et al., 1994; Lorenz et al., 1996; McDermid et al., 1996; Granot et al., 2001; Crombez et al., 2004). Hypervigilance is marked by increased attention and cognitive processing of a variety of external and internal noxious sensations and involves amplification of aversive adverse events (McDermid et al., 1996). In this case, hypervigilance would be marked by higher scores to the different situations of the SPS. However, a second potential explanation is provided by the existence of substantial and growing evidence supporting the mechanism of neurophysiological disturbances in fibromyalgia (for a review, see Vierck, 2006). Among other things, these disturbances include central sensitization induced by peripheral inputs from deep tissue and from cutaneous nociceptors, magnifying pain (Kosek et al., 1996). This sensitization is generalized across the body and across the physical nature of the stressor (mechanical pressure or thermal stimulus). Consequently, it could explain the higher SPS measures found in the fibromyalgia group as all situations, regardless of the body region involved, would be perceived as more painful compared with other pathologies such as low back pain, for which pain is more localized so that only situations involving the affected region would appear more painful. Those situations, for instance 'I wrench my back lifting a heavy weight', presented a different relative painfulness for the low back pain subgroup compared with the other subgroups (that is, a DIF), and were deleted from the original 43-item questionnaire.

A first variable significantly correlated with the SPS is PPI. One possible explanation for this correlation is that the current pain status influences the representation of the different situations in the sense that when the PPI is high, chronic pain patients rate the situations as being more painful and *vice versa*. Current pain status has already been described as influencing pain recall. When the PPI was high, chronic pain patients rated their levels of prior pain as being more severe than previously indicated (Eich et al., 1985; Feine et al., 1998). An alternative explanation for this correlation is that the SPS effectively reflects the attitude of the subject and this attitude influences the PPI in the same direction.

A second variable significantly correlated with the SPS measure is pain-related anxiety (PASS total and subscale scores). Anxiety specifically related to pain can cause an increased pain salience exaggerating its aversiveness and leading to higher pain perception reports. The SPS measure correlates more significantly with the escape-avoidance and the physiological anxiety subscales. The association with the escape-avoidance subscale can be explained by the fact that beliefs and memories of past painful experiences influence both avoidance and pain representation. Association with the self-reported physiological reactions may reflect a common bias in attentional and interpretational processes, for example, a tendency to amplify pain-related responses, as suggested by research on hypochondria and hypervigilance (Chapman, 1986). Contrary to the PASS, other psychological variables, such as the levels of depression (HAD-D) or anxiety (HAD-A and STAI-T), do not correlate with the reported situational pain representation. This fact suggests that anxiety specifically related to pain is more relevant to explain pain report than are more general tendencies to anxiety or depression.

In conclusion, the SPS has been developed and validated with a powerful psychometric tool, the Rasch model, leading to a common scale for both healthy subjects and chronic pain patients. Representation of imaginary painful situations differs greatly from one subject to another, in the healthy as well as in the chronic pain population. Anxiety about pain and PPI were found to be significantly correlated with the representation of imaginary painful situations. Future studies are needed to investigate the capacity of this instrument to predict a patient's behavior such as coping strategies developed by the pain sufferer or the risk of chronification for a patient in acute pain. In this regard, although the correlation methodology of the present study does not permit causal explanations, associations between the SPS and escape and avoidance behaviors (PASS-EA) can be considered promising. Another point to study is the evolution of this pain representation between an acute and chronic pain condition. Association between the SPS and social variables, such as level of education or household income, should also be examined to further delineate factors associated with the representation of pain.

Chapter 3
Investigation of rating scale functioning

3.1. Theoretical considerations

The problem of the optimal number of response categories in general

When an item presents more than two response categories, its response format is called polytomous. A polytomous response format is supposed to provide more sensitive information than would be acquired from a simple dichotomy. In theory, this sensitivity should increase as the number of response categories increases, pleading for finely graduated response scales. Nevertheless, there is probably an upper limit to the number of response levels that can be distinguished when assessing a latent variable (for example, 101 categories is unlikely to be more sensitive than 11 response categories when asking a subject to assess a particular latent attribute). Beyond a certain number of categories, confusion appears in the respondents' minds. As a result, the addition of response categories, instead of improving precision, decreases the accuracy of the response because these additional categories introduce error. As rating scales are widely used, determination of the optimal number of categories becomes an important consideration in the construction of such scales.

A lot of research literature has examined the question of the most appropriate number of response categories for the optimal measurement of a given variable. Traditionally, the method to solve this problem has been to determine the effect of number of rating categories on the reliability of the responses. Results led to divergent conclusions. Bending (1954a), Komorita and Graham (1965), and Matell and Jacoby (1971) found that reliability is generally independent of the number of response categories. Finn (1972) and Ramsay (1973) showed that the optimum number turn out to be seven confirming the hypothesis of Symonds (1924) and the assertion of Miller (1956). Miller (1956) examined a variety of published and unpublished research involving absolute judgment of simple, unidimensional stimuli (e.g., judgment of tones, loudness or taste intensity). He concluded that there is a clear and definite limit to the accuracy with which subjects can identify absolutely the magnitude of such stimuli, and that this span of absolute judgment is in the neighborhood of seven classes or alternatives. Lissitz and Green (1975) and Jenkins and Taber (1977), simulating item scores, showed a definitive leveling off in the increase in reliability after 5 categories. The study of Bending (1954b) indicated that rater reliability was highest with a 4-category scale. Reconciliation of these results is difficult because of the variety of variables being rated and of methodological approaches applied across studies. A lot of them emphasize reliability as the major, and in some instances, only criterion in the choice of number of scale categories. However, according to Matell and Jacoby (1971), the ultimate criterion should be the effect a change in the number of scale categories has on the validity of the scale. Their study investigated the effect of variations in the number of response levels (from 2 to 19) on the reliability and on the validity of the response scale. They concluded that both reliability and validity were independent of the number of response levels.

The optimal number of response categories for the assessment of pain

For the assessment of pain intensity, the gold standard is the Visual Analog Scale (VAS) usually scored from 0 to 100, i.e., on 101 response levels. Other most commonly used methods to assess pain intensity are the Numerical Rating Scale (NRS) that generally proposes 11 or 101 numerical response levels, and the Verbal Rating Scale (VRS) consisting of 4 (e.g., Breivik et al., 2000) to 15 (Gracely et al., 1978) adjectives depicting pain intensity. As can be seen, the number of pain intensity levels varies considerably from scale to scale. Clinicians and researchers often prefer VASs or NRS-101 for their high number of response categories. Because VRSs provide fewer response categories, they are usually assumed to be less sensitive. However, even though pain scales with more response categories have the potential to be more sensitive, such scales are not necessarily more sensitive or valid by virtue of the number of response categories alone. Some authors have already highlighted the existence of an upper limit to the number of response categories necessary to assess pain intensity. In laboratory conditions, Hardy and coworkers (1952) showed that healthy subjects are able to identify 21 just noticeable differences between the pain threshold and the level of pain intensity beyond which it becomes impossible to distinguish between stimuli (intolerable pain) meaning that scale sensitivity is likely to be maximal with 22 levels. However, this result should be moderated because subjects were each of the three authors. We may suppose that they were extremely familiar with the stimulus and thus particularly well-trained at distinguishing among various intensities. Jensen and coworkers (1994) provided an empirical evaluation of the number of response levels needed to assess pain intensity in chronic pain patients by recoding 101-point NRS data to 2-, 3-, 4-, 6-, 11-, and 21-point scales. Correlation coefficients were calculated between scores on the original 101-point NRS and on each recoded scale. Results indicated that this coefficient was greater than 0.99 for the 11- and 21-point scales but began to decrease for the 6-point scale and were weakest for the 2-point scale. Furthermore, pre- to post-treatment changes were examined for the 101-point NRS and for each of the six recoded scales. Sensitivity to change was essentially the same for the 6-, 11-, 21-, and 101-point scales. Finally, an examination of the actual responses to the 101-point NRS showed that almost all patients used multiples of 5 in rating pain, effectively reducing it to a 21-point scale. They conclude that the 11- and 21-point scales contain, essentially, the same information as the 101-point NRS. Information began to slightly decrease with the 6-point scale. Two limitations of this study must be noted. The first limitation is that the pain intensity scores were recoded from a 101-point scale. Results might have been different if patients had rated their pain on different scales varying in the number of response levels. The second limitation is the application of parametrical statistics (e.g., paired *t* tests for the sensitivity to change) on ordinal scores. Hunter and coworkers (2000) and Pesudovs and Noble (2005) investigated the psychometric characteristics of 7-point faces scales in children and in adults respectively. Both studies showed confusion in the middle of the scale and suggest reducing to six (Hunter et al., 2000) or to five or six (Pesudovs and Noble, 2005) the number of faces. Nevertheless, none of them empirically tested the suggested scale.

The Rasch methodology as a tool for investigating the functioning of rating scales

Probably, there is no single optimal number of response categories that applies to all rating situations. The optimum number of rating categories should be determined empirically every time a new rating scale is developed or when an existing rating scale is used with a new population (Bond and Fox, 2001). Indeed, the number of response levels beyond which the subject cannot distinguish anymore is clearly function of the variable being rated and of the subject population. The Rasch methodology gives an effective framework for testing rating scale assumptions. It provides statistics that guide the analyst in the assessment of the functioning of response categories (Linacre, 1995; Bond and Fox, 2001; Linacre, 2002). The 'functioning' of response categories may be defined as the manner in which categories cooperate to provide the measure. Statistics provided by the Rasch analysis are described below and illustrated with two data sets. Both data sets are ratings of items of the Situational Pain Scale whose development and validation has been presented in chapter 2. Data of the first set are ratings of 100 healthy subjects on a four-level pain scale: not painful (0), slightly painful (1), moderately painful (2) or extremely painful (3). This first data set comes from the study presented in chapter 2 and is an example of a well-behaved rating scale. Data of the second set are those used in the study of section 3.2., that is ratings of 102 healthy subjects on a 10-level pain scale: indifferent (0), slightly cumbersome (1), unpleasant (2), very slightly painful (3), slightly painful (4), painful (5), clearly painful (6), extremely painful (7), atrocious pain (8) and the worst pain as possible (9). This second data set is an example of a problematic rating scale. Both data sets were analysed with a rating scale model.

The rating scale model (Andrich, 1978b) is a model that can be applied to the analysis of polytomous response formats. The specificity of this model resides in the fact that it forces all items to adopt the same structure of the response scale. A schematic representation of the response scale of three items analyzed according to a rating scale model is shown in the upper part of Figure 1. Thresholds, which are in fact the boundaries between the categories, have the same relative location. As a result, the rating scale analysis provides category statistics that apply to the entire set of items. The detail of the statistical indexes can be summarized in one table, which is particularly appropriate for the illustrative purpose of the present theoretical section. In contrast to the rating scale model, in the partial credit model (Wright and Masters, 1982), each item is free to adopt its own structure of the response scale. A schematic representation of the response scale of three items analyzed according to a partial credit model is shown in the bottom part of Figure 1. The partial credit analysis provides a set of individual threshold estimates for each item and, consequently, category statistics for each item individually allowing a much more accurate and specific analysis. For this reason, the partial credit model has been retained for the investigation of the category functioning of the response scales in the empirical studies of the following sections of this chapter.

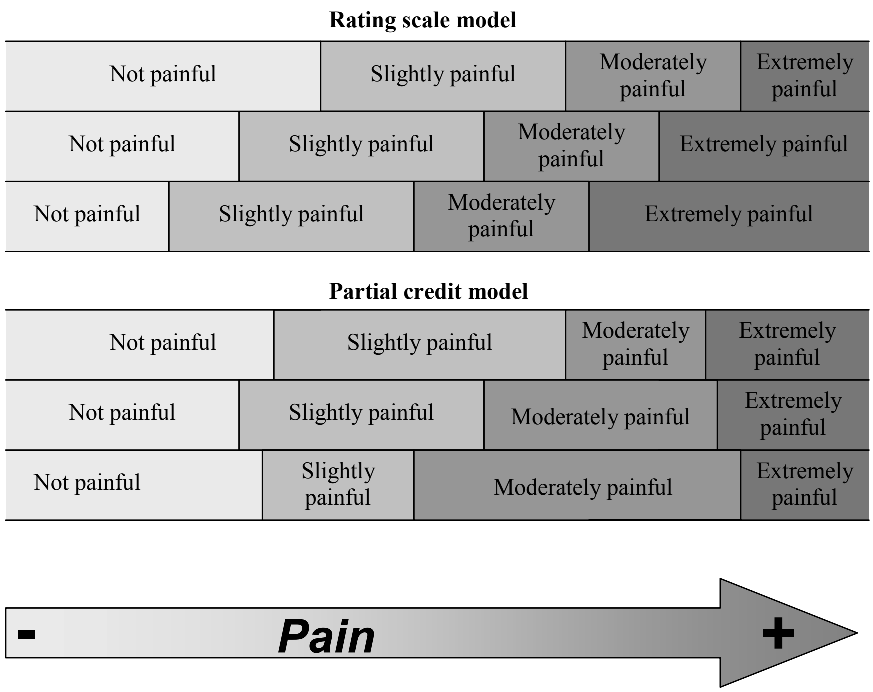


Figure 1. Schematic representation of the response scale of three items analyzed according to a rating scale model (upper panel) and according to a partial credit model (bottom panel).

1) Category frequencies

The analysis of the category frequencies is the simplest way to assess category functioning. Two elements must be examined: the number of responses in each response category and the shape of the distribution of responses across categories. A sufficient number of responses per category is essential to obtain stable threshold estimations. Thresholds are the boundaries between categories. Linacre (2002) recommends a minimal number of ten observations in each category. Below this limit, including or excluding one observation considerably modifies the threshold estimation. Categories that are chosen less frequently often indicate that they are unnecessary or redundant. The second element that must be taken into account is the form of the distribution of ratings across categories. A uniform distribution is optimal because it indicates a regular category usage. Unimodal, bimodal or slightly skewed distributions are satisfactory. Irregular distributions such as highly skewed or 'roller-coaster' distributions are problematic because they indicate irregular category usage.

In the example of the 4-level rating scale, the number of response in each category is high and the distribution is unimodal peaking in central categories (Table 1, column 2 and 3). In the 10-level rating scale, low frequencies are observed for the highest categories and the distribution is irregular (Table 2, column 2 and 3).

Table 1. Category statistics for the four-level pain rating scale

Category label	Category count	Category %	Average measure	Mnsq fit statistics		Threshold calibration	Coherence %	
				infit	outfit		M→C	C→M
0	544	15	-2.18	1.03	1.02		70	35
1	1449	40	-0.80	0.96	0.95	-2.48	60	74
2	1088	30	0.74	0.97	0.98	0.24	53	60
3	583	16	2.63	1.04	1.05	2.24	76	49

Table 2. Category statistics for the ten-level pain rating scale

Category label	Category count	Category %	Average measure	Mnsq fit statistics infit	outfit	Threshold calibration	Coherence % M→C	C→M
0	113	6	-1.45	1.42	1.33		100	1
1	233	13	-1.29	0.99	0.99	-2.16	32	21
2	428	24	-1.11	0.82	0.86	-1.78	42	44
3	222	12	-0.80	0.79	0.74	-0.26	16	34
4	249	14	-0.57	0.76	0.71	-0.79	26	38
5	314	17	-0.26	0.92	0.95	-0.67	37	27
6	141	8	0.01	0.92	0.96	0.63	30	27
7	83	5	0.34	0.96	0.99	0.64	47	10
8	21	1	0.41	1.37	1.26	1.76	0	0
9	3	0	-0.61	3.65	3.91	2.63	0	0

2) Average measures

The observed response is the score directly obtained from the response of the subject on the rating scale; for instance, the observed response of a subject choosing 'slightly painful' on the 4-level verbal rating scale is 1. This observed response is function of the difference between subject's ability and item difficulty. In the context of the present thesis, the subject ability is viewed as the subject's pain perception and the item difficulty as the item painfulness. The average measure for a particular category is defined as the average of this difference between the subject's perception and the item painfulness across all observed responses in that category. As, usually, subjects with higher pain perceptions respond in higher categories, average measures are expected to increase monotonically with the rating scale. Lack of monotonicity in the average measures indicates that the categories do not function in the intended order.

In Table 1, the four-level rating scale shows a monotonic progression in the average measures from –2.18 to 2.63 logits. In Table 2, the average measure corresponding to the three observations of the last category fails to advance. Moreover, average measures of the other categories barely increase.

3) Fit statistics

The stochastic nature of the Rasch model implies a certain amount of randomness throughout the data. Too little randomness signifies that data are too predictable; excessive randomness is a sign of too much 'noise' in the data. The amount of randomness is modeled by mean-square fit statistics. Computational details for this index are presented in Linacre (1995). The expected value of randomness is indicated by a mean-square of 1.00. Values over 2.00 suggest that there is more unexpected than expected randomness. Such values indicate that this category has been used in contexts in which the expected category is far different.

Table 1 shows that that the fit of each category is very close to its expected value of 1.00. In Table 2, category 9 has mean-square (mnsq) values above 3.00 signifying that there is much more noise than expected by the stochastic nature of the model. The three observations in that category are highly unpredictable. Category 0 and 8 are also somewhat noisy. This example illustrates the fact that "unexpected use of an extreme category is more likely to produce a high mean-square than unexpected use of a central category" (Linacre, 2002). Central categories often exhibit over-predictability, that is mean-square fit statistics lower than 1, especially in situations where respondents are cautious or apathetic.

4) Thresholds

A threshold is the location along the continuum corresponding to an equal probability of observing two adjacent categories. Thresholds define the boundaries between the categories. Therefore, like the average measures, thresholds locations should increase monotonically (Andrich et al., 1997; Linacre, 1999). Thresholds that do not increase monotonically are considered disordered. Additionally, the distance between the threshold estimates is also critical. Ideally, it should be comprised between 1.4 and 5 logits (Linacre, 2002). A distance of at least 1.4 logits[a] is necessary in order for the rating scale of m+1 categories to be equivalent to a sub-test of m dichotomies, that is for the different categories to be clearly distinct. When adjacent thresholds are more than 5 logits far apart, then the category they define becomes too wide and the measurement loses its precision. In Table 1, thresholds are ordered and sufficiently distant from each other (2.72 and 2 logits). The 10-level rating scale, on the other hand, presents several disordered thresholds (Table 2).

[a] This critical value applies to three-category scales and lessens as the number of categories increases (Linacre, 2002).

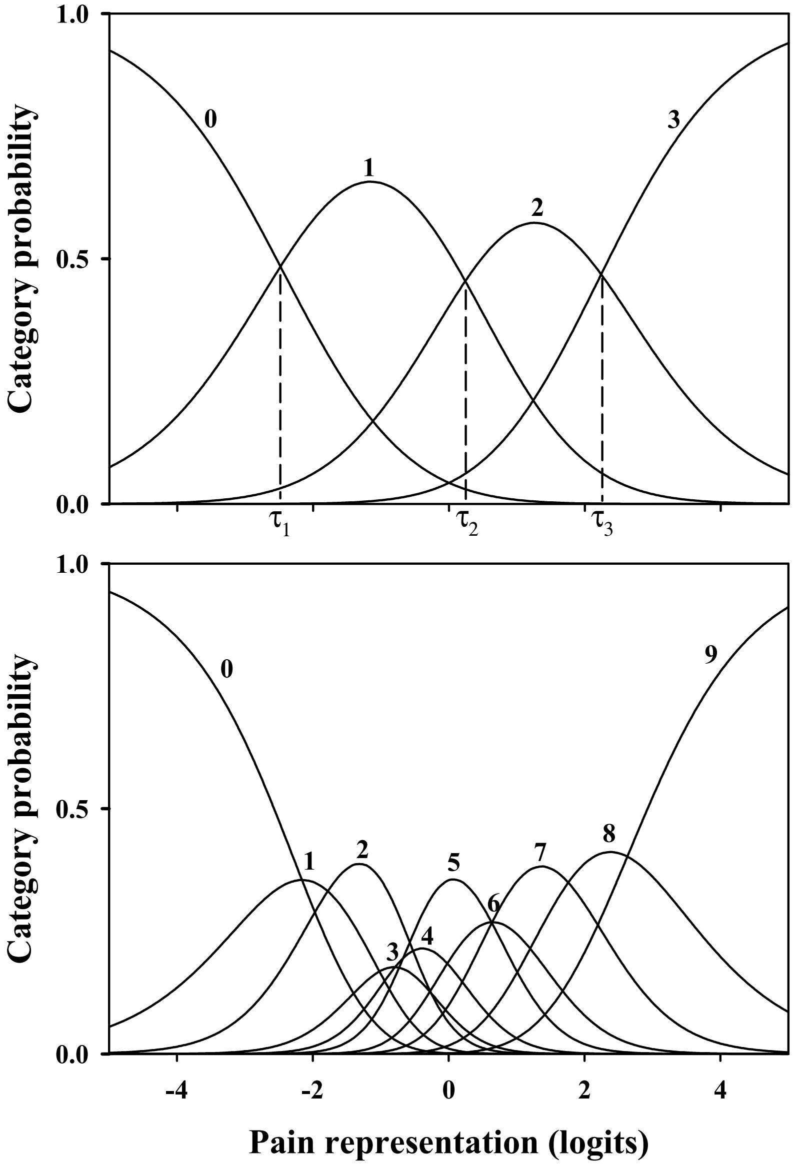


Figure 2. Category probability curves for the 4-level (top panel) and the 10-level (bottom panel) response scale indicating the probability of each response category as a function of the subject's representation of pain. Thresholds (τ) are located at the pain representation level for which pairs of successive response are equally probable (shown on top panel). For the ten-level response scale, at no point are categories 3 and 4 the most likely to be selected (bottom panel). Additionally, limits of category 6 (τ_6 and τ_7) are confounded.

The criterion of ordered thresholds can be visualized with the category probability curves (CPC). Category probability curves present the probability of each response category in function of the pain representation. The top panel of Figure 2 displays the CPC for the data of Table 1. Thresholds, symbolized by a τ, correspond to the projection, on the x-axis, of intersections of successive category probability curves. These probability curves look like a range of hills meaning that each category in turn is the most probable and that thresholds advance monotonically. The bottom panel of Figure 2 displays the CPC for the data of Table 2. Categories 3 and 4 are never the most likely to be observed; thresholds 3, 4 and 5 are disordered. Category 6 barely emerges but its limits (τ_6 and τ_7) are confounded (Table 2).

5) Coherence between observed and expected scores

On the basis of the estimated measures of subjects, items and thresholds, the model can compute an expected score for each subject-item interaction. The coherence reflects the degree of consistency between observed and expected scores. Two coherence indexes can be computed. The first index, noted 'M→C' (Measure implies Category), represents the percentage of scores expected to be observed in a category that are actually observed in that category. The second index, noted 'C→M' (Category implies Measure), represents the percentage of scores observed in a category that are expected in that category by the model (Linacre, 2002). Linacre (2002) recommends a minimal level of coherence of 40%. According to the user's guide to Winsteps, a Rasch-model computer program, levels of coherence lower than 50% are inferential insecure (Linacre and Wright, 1998).

In Table 1, majority of coherence levels are above these limits. Only the 'C→M' index of category 0 is poor. No more than 35% of scores observed in category 0 are expected in that category by the model. On the other hand, in Table 2, almost all coherence levels are below the critical value of 40%. Only category 2 presents acceptable levels of coherence between observed and expected scores.

To sum up, Rasch analysis provides several useful indicators to investigate rating scales categorization. These different indexes are often redundant and should be used in combination. Some, such as the order of thresholds and the levels of coherence are stricter. Most used diagnostic statistics to investigate the rating scale functioning are the category frequencies, the average measures and the thresholds (Zhu et al., 1997; Arnould et al., 2004; Pesudovs and Noble, 2005). Fit statistics and coherence levels are less employed.

Remedies for improving the functioning of rating scales

When rating scale investigative statistics reveal that the categories fail to be used as intended, several remedies exist to improve the functioning of the rating scale and subsequently, the measurement accuracy. The first solution consists in renumbering the categories in case of obvious reversed categories as indicated by disorder in the average measures. Practically, this situation is quite infrequent as categories should normally be carefully defined and labeled. The second solution is to collapse adjacent categories when some categories are used infrequently (indicated by low category frequencies) or inconsistently (indicated by disordered thresholds or low levels of coherence). This is the most frequent scenario. Third solution is to drop items if problems in the rating scale only concern a few items, meaning that the problem probably resides in these items rather than in the rating scale. For example, these items may assess multiple dimensions. The last and the least frequent solution consists in adding categories. It only applies when distance between thresholds is larger than 5 logits. Except for this last solution, once the rating scale has been modified, data must be reanalyzed. Practically, often, several combinations must be tested to find the optimal categorization in terms of category functioning and reliability of the scale. Andrich (1996) noted that when thresholds are correctly ordered and data fit the model, no further categories should be combined. If they were combined, data would fit the model worse. The study presented in section 3.2. will show the effects of grouping categories on the functioning of a response scale. These combinations are interesting because they can improve the interpretability of the resulting measures. Of course, these *a posteriori* combinations are somewhat artificial in the sense that we cannot be sure that respondents would use the new rating scale as postulated by the collapsing of categories. This grouping procedure gives an idea of the maximum number of levels a sample can distinguish and helps the investigator to improve his scale for future investigations. Although Zhu (2002) showed that the characteristics of the optimal categorization identified by the Rasch post-hoc analysis in a previous study (Zhu et al., 2001) could be maintained when the revised scale was applied to the same populations, each new combination must be validated empirically.

Specific aims of the next sections

The two next sections of this chapter will investigate the question concerning the optimal number of response categories for the pain intensity attribute. The first section is devoted to the particular case of verbal rating scales in a healthy adult population (section 3.2.) and the second one to faces scales in healthy children aged from four to seven years (section 3.3.). The common point of these studies is the application of the Rasch methodology for testing rating scale assumptions.

3.2. Assessment of pain intensity in adults using verbal rating scales. What is the optimal number of response levels?

Abstract

The verbal rating scale (VRS) is one of the most common scales used to assess pain intensity in adults. Many different VRS lists have been created differing among other criteria in the number of pain intensity levels. The objective of this study is to investigate the category functioning of three VRSs (four, five and ten response categories) in healthy adults in order to determine the most optimal number of response levels. Healthy adults rated the pain they would experience for 18 imaginary painful situations on one or several of the aforementioned verbal rating scales. Investigation of the category functioning was performed with the Rasch model. Results showed that the 10- and the 5-level VRSs were not effective to assess pain intensity but that the 4-level one was. Recategorization procedures applied on the 10- and the 5-level VRSs revealed that the optimal number of response levels was also around four. In conclusion, this study shows that the human being can only distinguish four response levels when assessing the pain intensity of hypothetical events on a VRS.

1. Introduction

The verbal rating scale (VRS) is one of the most common scales used to assess pain intensity in adults (Jensen and Karoly, 1992). They consist of a list of adjectives describing different levels of pain. Patients are asked to select the word or phrase that best describes their level of pain. They are easy to administer and to score. Because they are generally easy to comprehend, compliance rates for VRSs are as good or better than those for other measures of pain intensity under most conditions (Jensen et al., 1986; Jensen et al., 1989). Verbal rating scales have demonstrated their validity as indicants of pain intensity. They are positively and significantly related to other measures of pain intensity (e.g., Jensen et al., 1986; Jensen et al., 1989; Kremer et al., 1981; Littman, 1985). Some studies have reported a lower sensitivity of VRS compared with the visual analog scale or the numeric rating scale (Breivik et al., 2000; Wallenstein et al., 1980) but this lowered sensitivity has often been assumed as VRSs provide fewer response categories (Huskisson, 1974; Jensen et al., 1986). Many different VRS lists have been created (Gracely et al., 1978; Melzack, 1975; Breivik et al., 2000) differing among other criteria in the number of pain intensity levels.

The aim of the present study is to investigate, with the Rasch model (Rasch, 1960), the functioning of two VRSs, a 5-level one and a 10-level one in healthy adults in order to determine the most optimal number of response levels. The generalization of the study was extended by including data from the study presented in chapter 2 where a 4-level VRS was used.

2. Subjects and methods

2.1. Subjects

One hundred and two healthy subjects (41 males and 61 females) aged between 18 and 65 years (mean: 37, SD: 14.5) participated in this study. All subjects were native French speakings.

2.2. Instrument and procedure

Subjects assessed the pain they would experience for each of the 18 situations of the Situational Pain Scale (SPS) (chapter 2, Table 1) on two different VRSs. The first VRS comprised ten levels: indifferent (0), unpleasant (1), slightly cumbersome (2), very slightly painful (3), slightly painful (4), painful (5), clearly painful (6), extremely painful (7), atrocious pain (8) and the worst pain as possible (9); the second VRS comprised five levels: not painful (0), slight pain (1), moderate pain (2), severe pain (3) and extreme pain (4).[a] The two scales were presented in random order. Data from chapter 2 were added to the analysis. In that study, 100 healthy subjects and 111 chronic pain patients rated the same items on a 4-level VRS: not painful (0), slightly painful (1), moderately painful (2), and extremely painful (3).[b] Only ratings of healthy subjects (45 males and 55 females; aged between 20 and 59 years, mean: 39, SD: 11.2) were retained for the present study.

2.3. Rasch analysis

Subjects' responses on the ten-, five- and four-level VRSs were analyzed separately with the Winsteps© Rasch analysis software (Linacre and Wright, 1998) using a partial credit model (Wright and Masters, 1982). The partial credit model is an extension of Rasch's original dichotomous model (e.g., painful/not painful) (Rasch, 1960) for polytomous response formats, i.e. for items graded on more than two categories. The partial credit model allows each item to vary in its structure of the response scale. The estimation of the thresholds, which are the boundaries between the categories, differs for each item. As a consequence, the category functioning is investigated for each individual item.

Investigation of category functioning

Among the different statistical indexes presented in section 3.1., the category frequencies and the monotonic progression of average measures and thresholds were retained in order to investigate the category functioning.

[a] The scales were presented in French with the following category labels: 'indifférent' (0), 'désagréable' (1), 'légèrement gênant' (2), 'très légèrement douloureux' (3), 'légèrement douloureux' (4), 'douloureux' (5), 'nettement douloureux' (6), 'extrêmement douloureux' (7), 'douleur atroce' (8) and 'la pire des douleurs' (9) for the ten-level VRS; 'pas douloureux' (0), 'douleur légère' (1), 'douleur modérée' (2), 'douleur sévère' (3) and 'douleur extrême' (4) for the five-level VRS.

[b] French labels of that scale were: 'pas douloureux' (0), 'légèrement douloureux' (1), 'moyennement douloureux' (2) and 'extrêmement douloureux' (3).

Improvement of category functioning

When average measures and/or thresholds fail to progress monotonically, they reveal problems in the use of the different categories. Then, adjacent categories were collapsed and data were reanalysed. The aim of this procedure is to improve the functioning of the rating scale and subsequently the interpretation of the measure. There are no fixed rules for collapsing categories, but only guidelines (Wright and Linacre, 1992; Linacre, 2002). The first and foremost guideline in collapsing rating scale is that what we collapse must make sense. For example, it does not make sense to collapse a category labeled 'not painful' (below the pain threshold) with a category labeled 'slightly painful' (above the pain threshold). Then, categories to collapse are generally those presenting lower usage frequencies or average measures and/or thresholds that do not progress monotonically. Numerous collapsings were explored varying in the number of response levels and in the choice of categories that are collapsed. When comparing these categorizations, we will examine the average measure and threshold progression for each item but also the separation reliabilities for items and subjects. The separation reliability coefficient is determined as the ratio between the true measure variance and the observed (true + error) measure variance (Wright and Masters, 1982; Fisher, 1992; Wright, 1996). It corresponds to the proportion of observed variance that is not due to measurement error but to real variation in the measure. In other words, this index indicates the reliability with which measures can be separated, with values ranging between 0 and 1 (Wright and Masters, 1982). This index can be computed for item and subject measures. High (near 1.0) separation reliability signifies that the variance is due to true measure variance (not to measurement error variance) and is preferred.

3. Results

3.1. The 10-level verbal rating scale (VRS (0-9))

Investigation of the VRS (0-9)

Results of the partial credit analysis are reported in Table 1. The analysis revealed that only one item had ordered average measures and another item ordered thresholds. All the other items presented multiple average measure and threshold disorders. The two highest response categories had very low category frequencies: 21 responses in category 'atrocious pain' and 3 responses in category 'the worst pain possible', all items and all subjects taken together. These results show that this 10-level VRS is not efficient in assessing the pain intensity of imaginary painful situations.

Table 1. Category functioning of original and recoded scales.

	VRS (0-9)	Grouping a	Grouping b	VRS (0-4)	Grouping c	VRS (0-3)
Number of response levels	10	4	3	5	4	4
Number of items with ordered average measures	1	13	14	11	16	17
Number of items with ordered thresholds	1	17	18	14	17	18
Person separation reliability	0.88	0.86	0.86	0.85	0.85	0.90
Item separation reliability	0.93	0.97	0.97	0.94	0.97	0.98

Rescoring of the VRS (0-9)

Given the poor functioning of the original VRS (0-9), adjacent categories were collapsed in order to produce rating scales that yield higher quality measures. Several categorizations were explored. Only results of categorizations producing the best category functioning are presented. When the original ten categories were collapsed into five or more categories, more than half of the items had disordered average measures and disordered thresholds, whatever the grouping. The most satisfying reduction into a four-level scale is displayed in Figure 1 (grouping a) and was obtained by grouping category ‘indifferent’ with ‘unpleasant’; category ‘slightly cumbersome’ with ‘very slightly painful’; category ‘slightly painful’ with ‘painful’; and category ‘clearly painful’ with ‘extremely painful’, ‘atrocious painful’ and ‘the worst pain possible’. Analysis of these rescored data resulted in a substantial improvement of the number of items with ordered average measures (13 out of 18) and with ordered thresholds (17 out of 18) (Table 1). Person separation reliability index slightly decreased from 0.88 to 0.86 while item separation reliability index raised from 0.93 to 0.97. The most satisfying reduction into a 3-level scale is displayed in Figure 1 (grouping b) and was obtained by grouping category ‘indifferent’ with ‘unpleasant’ and ‘slightly cumbersome’; category ‘very slightly painful’ with ‘slightly painful’ and ‘painful’; and category ‘clearly painful’ with ‘extremely painful’, ‘atrocious painful’ and ‘the worst pain possible’. Compared with the four-level recategorization, this three-level one very slightly enhances the number of items with ordered average measures (14 out of 18) and with ordered thresholds (18 out of 18). Separation reliability indexes are unchanged. Nevertheless, in terms of linguistics, the three-level grouping makes more sense as the new pivot point between category ‘slightly cumbersome’ and ‘very slightly painful’ corresponds to the pain threshold. Further reductions to a dichotomous response scale led to important decreases in person and item separation indexes.

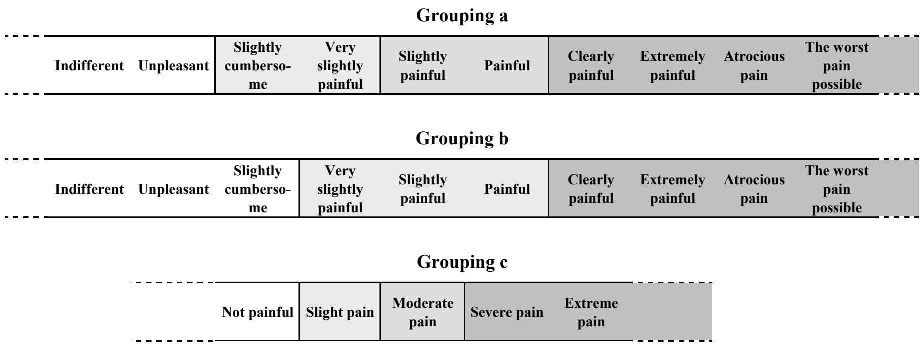


Figure 1. Grouping a: most satisfying collapsing of the original 10-level verbal rating scale into a 4-level scale. Grouping b: most satisfying collapsing of the original 10-level verbal rating scale into a 3-level scale. Grouping c: most satisfying collapsing of the original 5-level verbal rating scale into a 4-level scale.

3.2. The 5-level verbal rating scale (VRS (0-4))

Investigation of the VRS (0-4)

The partial credit analysis of the 18 items rated on the VRS (0-4) showed that 11 items had ordered average measures and 14 items had ordered thresholds (Table 1). The highest response category was chosen for only 2% of the ratings. Person and item separation indexes equal 0.85 and 0.94 respectively. Taken into account that items with ordered average measures are not automatically the same than those with ordered thresholds, this scale was satisfying for less than 50% of the items.

Rescoring of the VRS (0-4)

Given that for the great majority of items, problems in the functioning of the VRS (0-4) came from the lack of responses in category 'extreme pain', this category was collapsed downward into category 'severe pain' (grouping c, Figure 1). This grouping resulted in five additional items presenting ordered average measures and in three additional items presenting ordered thresholds. Item separation reliability rose from 0.94 to 0.97; person separation reliability remained unchanged (Table 1). Further groupings into three or less categories only slightly improved the number of items with ordered average measures and the number of items with ordered thresholds. It is not surprising given that these criteria were already almost optimal with the reduction into a 4-level VRS. On the other hand, these further groupings importantly lowered separation reliability indexes.

3.3. The 4-level verbal rating scale (VRS (0-3))

The partial credit analysis of the 18 items rated on the VRS (0-3) led to almost perfect results in terms of number of items with ordered average measures and ordered thresholds. Separation reliability indexes have the highest values compared with all other original scales and groupings (Table1). Given these results, no groupings were performed with this scale.

4. Summary and conclusion

When healthy adult subjects rated items of the SPS, the functioning of the ten- and five-level VRSs was not satisfying. Among the scales presented in their original format, the 4-level VRS 'not painful', 'slightly painful', 'moderately painful' and 'extremely painful' produced the best category functioning and the highest indexes of separation reliability. It must be reminded here that the sample who used the 4-level VRS was not the same as the one who rated the 18 situations on the 5- and the 10-level VRS. Nevertheless, demographical characteristics of both samples were very similar so that there is no reason to think that it could explain these results. The effectiveness of the 10-level VRS was optimized with the categories grouped into three levels. The recategorization of the 10-level VRS into four levels produced almost as satisfactory results as the grouping into three categories. Concerning the 5-level VRS, its effectiveness was optimized when categories were grouped into a 4-level scale. The concordance between recoded and original scales concerning the optimal number of response levels supports the hypothesis that the grouping and reanalyzing procedure can produce very good approximations of the optimal number of response categories and can be use to guide the improvement of existing scales.

In conclusion, in view of these results, a 4-level scale is recommended for the assessment of pain in adults, although these findings should be empirically confirmed for other labels, other types of scales (numerical rating scale for instance), other populations (patients) and other situations (experimental or clinical pain).

3.3. How many response levels do children distinguish on faces scales for pain assessment?

Abstract

Faces scales (FSs) are one of the most commonly used instruments to assess pain intensity in children. Most available FSs present five to seven faces. The present research was conducted to investigate the ability of 4- to 7-year-old children to distinguish the response categories of different FSs. In the first study, 121 children were asked to rate painful situations on a 6- and a 3-level FS (FS-6 and FS-3) commencing with a smiling 'no pain' face. Children were divided into two age groups (4-5 and 6-7 years). Investigations of the category functioning were performed with a partial credit Rasch model for each age group. Results revealed the low performances of the FS-6 as compared to the FS-3 and also the difficulty children experienced in scoring the imaginary painfulness of items. Consequently, a second study was conducted. In this second study, 76 children were asked to rate pictures depicting painful situations on a FS-3 beginning with a neutral 'no pain' face. Results of this second study confirmed an improvement in the ability to distinguish the three response categories with age. The 4-5 year-old children could only distinguish two response categories and the 6-7 year-old children were able to discern the three levels of the FS-3. In conclusion, young children do not distinguish as many faces as proposed by the majority of available faces scales. These results strongly recommend a reduction in the number of response levels of FSs for pain assessment in children.

Submitted for publication as:

Decruynaere C, Briart V, Taziaux N, Angelou G, Thonnard J-L, Plaghki L. How many response levels do children distinguish on faces scales for pain assessment.

1. Introduction

The faces scales (FSs) are self-report scales using pictures of facial expressions to assess pain intensity. They show a series of ordered faces conveying increasing intensity of pain between 'no pain' and 'worst pain possible' (for a review, see Chambers et al., 1999). Faces scales do not require the ability to count or to use numbers in an ordinal fashion. For this reason, they are often used with children. Many different FSs have been created varying among other criteria the number of faces. Most FSs have five to seven faces (Beyer, 1984; LeBaron and Zeltzer, 1984; Wong and Baker, 1988; Kuttner and Lepage, 1989; Bieri et al., 1990). This categorization of pain intensity variable is intended to elicit from children an ordinal indication about the intensity of pain. It is supposed to give more information than expected from a simple binary 'pain'/'no pain' response. However, whether children use FSs effectively and appropriately is still a matter of debate (Stinson et al., 2006).

The scale constructor may have given more categories than child can distinguish possibly leading to confusion in a child's mind. Some studies have already reported childrens' difficulty in differentiating between too many categories. Pesudovs and Noble (2005) analyzed adults' responses on a 7-category FS and showed improvement in category usage when reducing the number of categories to five or six faces. Hunter et al. (2000) reported that children aged from 3 to 6 years frequently confused response categories near the middle of a 7-level FS and proposed to reduce the number of faces to six.

An innovative approach to investigate a subject's capacity to distinguish the different categories on rating scales is the use of one member of the family of Rasch models, the partial credit model (Masters, 1982; Wright and Masters, 1982). The Rasch family of models belongs to a group of probabilistic models known as Item Response Theory. Usually applied to measure human performances, attitudes and perceptions (Tesio, 2003), the partial credit analysis also provides an effective framework within which to verify and improve the functioning of rating scale categorization (Linacre, 2002).

The present studies are aimed at examining with the partial credit model how a sample of healthy children aged from four to seven distinguish the different faces of FSs when rating imaginary painful situations. In the first study, children's ability to distinguish a 6- and a 3-level FS was tested. A 6-level FS (FS-6) was used because it corresponds to the average number of faces on the majority of available FSs (Chambers et al., 1999) and a 3-level FS (FS-3) because the aforementioned studies already emphasized the difficulty to differentiate many response categories. As results of the first study revealed the low performance of the FS-6 as compared to the FS-3 but also the difficulty children experienced in scoring the imaginary painful situations, a second study was conducted using a FS-3 and vignettes depicting painful situations in order to improve representation of the imaginary painful situations.

2. Study 1

2.1. Methods

2.1.1. Subjects

A total of 121 healthy children (89 boys and 32 girls) participated in this experiment. They ranged in age from four to seven years. They were recruited from various kindergartens, primary schools and sport centres in Belgium (French part of the country). After having received the authorization of the establishment, permission to assess children was obtained via written informed consent of the parents. The sample of children was divided into two age categories: 4 to 5 years (preschool), n = 72 (46 boys, 26 girls) and 6 to 7 years (primary school), n = 49 (43 boys, 6 girls) in order to examine age-related differences in the way children use of FSs. This stratification is consistent with school levels in Belgium and with the fact that children less than six years of age typically fail standard Piagetian tasks that their counterparts over six typically succeed in performing (Gelman and Baillargeon, 1983). The way children define pain and think about several aspects of pain has been shown to be consonant with the Piagetian developmental model (Gaffney and Dunne, 1986). No finer division into age groups were performed in order to keep sufficiently large groups for the subsequent Rasch analysis.

2.1.2. Faces scales

Two different FSs were presented to children: a 6-level FS (FS-6) and a 3-level (FS-3). They were similar in face design to those previously used in pain research (Chambers and Craig, 1998; Pesudovs and Noble, 2005) and conveyed increasing levels of pain intensity, from no pain to extreme pain (Figure 1, (a) and (b)). Faces scales lay on cards put on the desk in front of the child. Faces were 4 cm from top to bottom.

(a)

(b)

(c)

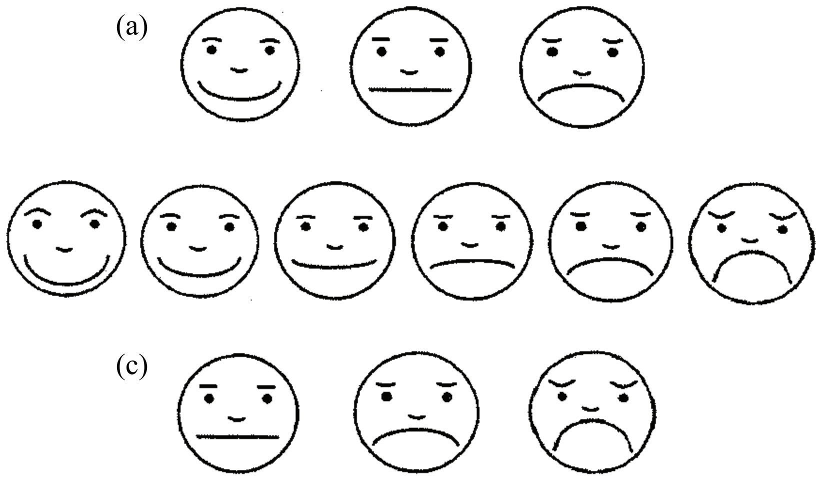


Figure 1. The faces scales. (a) The faces of the 3-level FS used in Study 1 from face 0 (no pain) on the left to face 2 (extreme pain) on the right. (b) The faces of the 6-level FS used in Study 1 from face 0 (no pain) on the left to face 5 (extreme pain) on the right. (c) The faces of the 3-level FS used in Study 2 from face 0 (no pain) on the left to face 2 (extreme pain) on the right.

2.1.3. Procedure

All participants were tested individually in a quiet room. The child was sitting comfortably at a desk opposite the experimenter. Before testing, a few minutes were spent in order to meet the child, making sure that he had correct notions of pain and to explain the task to be presented. Then, the child was described in French a series of 25 painful situations (called the items) developed for the purpose of the study (Table 1, upper list). Similar painful situations have already been presented to children of the same age to investigate the validity of pain scales (McGrath et al., 1996; Hunter et al., 2000). Situations were presented in random order. The child was asked to choose the face that best reflected the amount of pain he would feel in each situation. Children rated the entire list of situations twice, once on the FS-6 and once on the FS-3. The order of scale presentation was random. The entire session lasted approximately 40 minutes.

Table 1.

List of the 25 painful situations of Study 1	
1. A mosquito bit	14. Someone pulls the hair
2. Getting sunburned on the face	15. Getting the fingers caught in the car door
3. A toothache	16. Knocking the head on the corner of a piece of furniture
4. Walking barefoot on pebbles	17. Getting indigestion after a large meal
5. Cleaning a scratch	18. Catching fingers in a zipper
6. The doctor examine the throat	19. Getting sunburned and being touched on that spot
7. Being stung by a bee	20. The dentist examines the teeth in a check-up
8. Getting a speck of dust in the eye	21. Cutting the hand with a knife
9. Burning one's tongue tasting scorching hot food	22. Stubbing one's toe on a chair leg
10. Walking on burning sand	23. Biting one's tongue
11. Getting one's fingers frozen	24. Carrying a heavy bag
12. Getting shampoo in the eye	25. Spilling some boiling water on the hands
13. Getting an injection at the doctor's	
List of the 15 vignette situations of Study 2	
1. A mosquito bit	9. Getting indigestion after a large meal
2. Walking barefoot on pebbles	10. Cutting the hand with a knife
3. Cleaning a scratch	11. Someone stand on your foot
4. The doctor examine the throat	12. To lose a tooth
5. Getting shampoo in the eye	13. To put one's hand on a hotplate
6. Getting an injection at the doctor's	14. To take a bath
7. Someone pulls your hair	15. To play football (boys)
8. Getting the fingers caught in the car door	15. To skip with a rope (girls)

2.1.4. Rasch analysis

Responses were analysed with the Winsteps© Rasch analysis software (Linacre and Wright, 1998) using a partial credit model (Masters, 1982; Wright and Masters, 1982). The partial credit model is an extension of Rasch's original dichotomous model (e.g., painful/not painful) (Rasch, 1960) for polytomous response formats, i.e. for items graded on more than two response categories. For each child responding to each imaginary situation, the model formulates the probability of selecting each response category. By means of an iterative process, the partial credit model estimates a measure of pain representation for each child, the painfulness of each situation and the location of each threshold. Thresholds are defined as the location along the pain representation variable for which two adjacent response categories have an equal probability of occurrence and thus delineate the boundaries between the categories. In the partial credit model, threshold locations are free to vary from item to item. Each item has thus its own set of estimated thresholds. All these measures are located on the same scale and expressed in logits, the natural logarithm of an odds ratio. The left panel of Figure 2 displays the different concepts here above mentioned. The chart illustrates the category probability curves for a given situation rated on a 6-level scale. The x-axis represents the pain representation continuum and the y-axis the probability of each response category. Each curve corresponds to one response category. Subjects located on the left of the scale, with low pain representations, have category 0 as the most probable. As a subject's pain representation rises, each response category becomes in his turn the most probable. Thresholds correspond to the projection, on the x-axis, of intersections of successive category probability curves. Threshold one and four, symbolized τ_1 and τ_4 are shown as examples. Distance between two successive thresholds delimits the range for which one category is the most probable. The response scale can be represented like a ruler (bottom of Figure 2).

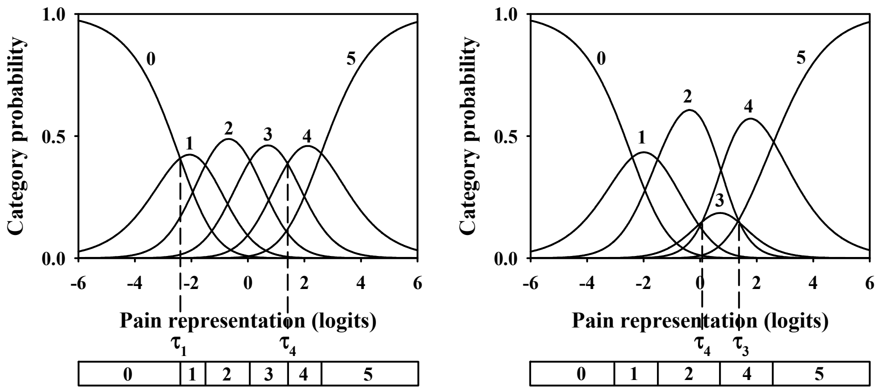


Figure 2. Category probability curves for a given situation indicating the probability of each response category as a function of the subject's representation of pain. Thresholds (τ) are located at the pain representation level for which pairs of successive response are equally probable. The bottom ruler is a synthetic representation of the response scale. Left panel: example of a well-functioning response scale. Thresholds 1 (τ_1) and 4 (τ_4) are displayed as example (broken lines). Right panel: example of a bad-functioning response scale. Category 3 never emerges. The probability of observing this category is lower than the probability of observing any other category whatever the subject's location along the scale. Thresholds delimiting categories 3 and 4 (τ_3) and 4 and 5 (τ_4) are reversed.

Several criteria have been proposed to investigate the functioning of rating scale categories (Linacre, 2002). The criterion retained for the present study is the monotonic progression of thresholds, implying that the category probability curves form a 'range of hills' such as illustrated in the left panel of Figure 2. An example of threshold disordering is illustrated in the right panel of Figure 2. The threshold between category 3 and 4 (τ_3) has a higher location along the pain representation scale than the threshold between category 4 and 5 (τ_4). Probability curves show that the fourth category is never the most likely to be selected. Reversed thresholds reveal that the categories are not functioning as intended.

2.2. Results

Separated Rasch analyses were conducted for each age group and for each FS. Distributions of child measures and item (or situation) locations for both scales are presented in Figure 3 (four left panels). Children reporting the highest pain scores are located on the right of the scale. Highest painful items are located on the left of the scale. Two major observations emerge. First, the range of item painfulness is extremely narrow for both scales and for both age groups reflecting the inability of children to differentiate situations according to their painfulness. For the FS-3, situation locations spread over 1.02 logit for the 4-5 year olds and over 0.93 logit for the 6-7 year olds. This effect is amplified with the FS-6: situation locations spread over 0.37 logit for the 4-5 year olds and over 0.60 logit for the 6-7 year olds. Second, children are located more on the right of the scale as compared to the item locations, reflecting children's tendency to favor high pain scores. This is confirmed by the percentage utilization of each response category, that is the percentage of occurrences of each category, presented in Table 2. Except for the first response category of the FS-6, both FSs present percentage utilization that increases together with the response categories. This phenomenon is greater for the younger age group than for the older one. Nevertheless, child's measures on both scales are correlated ($\rho = 0.70$, $p < 0.001$ for the 4-5 years and $\rho = 0.60$, $p < 0.001$ for the 6-7 years) revealing their rating consistency across scales.

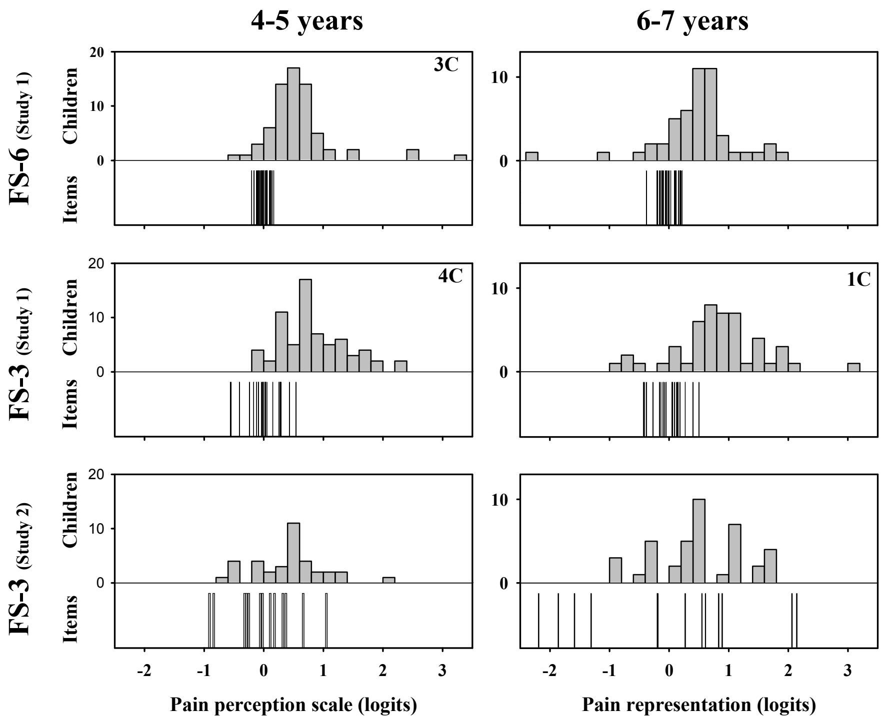


Figure 3. Distributions of children's measures and item locations estimated from responses on the 6-level FS (FS-6) of Study 1 (top panels), on the 3-level FS (FS-3) of Study 1 (middle panels) and on the 3-level FS (FS-3) of Study 2 at test time (bottom panels) for the 4-5 year-old (left panels) and for the 6-7 year-old children (right panels). Number of children presenting an extreme maximum score is marked on the upper right corner for each child distribution followed by a 'C' (children).

The examination of category functioning is summarized in Table 2. For the FS-6, all items present multiple threshold disorders regardless of the age category. On the other hand, the analysis of responses showed an improvement in the use of the FS-3 by age group. The number of items presenting ordered thresholds increases from 11/25 for the 4-5 year-olds to 18/25 for the 6-7 year-olds. Moreover, the mean distance between thresholds significantly increases with age (paired *t*-test, p = 0.002) (Figure 4). These results suggest that children of 4-5 years of age only distinguish two response levels and that children of 6-7 years of age begin to distinguish three response levels.

Table 2. Percentage utilization of categories and proportion of items with ordered thresholds

Age (years)	*N*	% utilization of categories									Proportion of items with ordered thresholds	
		FS-6						FS-3			FS-6	FS-3
		0	1	2	3	4	5	0	1	2		
4-5	72	11	8	12	14	17	38	19	26	55	0/25	11/25
6-7	49	12	9	15	18	18	28	18	32	50	0/25	18/25

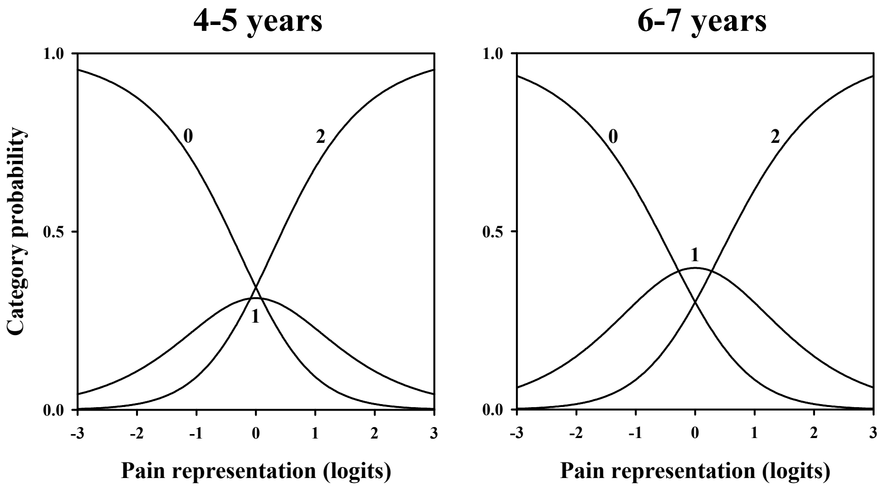


Figure 4. Category probability curves averaged across all items rated on the three-level faces scale for each age group.

3. Study 2

The first study revealed the low performances of the FS-6 as compared to the FS-3 and also the difficulty children experience in scoring the imaginary painfulness of items although they consistently rated the situations, not at random, as shown by the strong correlation between children's measures on both FSs. Taking these facts into consideration, a second study was conducted with the principal aim of enhancing children's performance in discriminating between items by modifying the test in four ways. First, we sought to clarify the representations of the portrayed painful situations by using vignettes depicting the imaginary events. Second, we sought to make the task simpler by reducing the number of items from 25 to 15 with the intent of shortening the task, reducing the burden and keeping attention focused. Third, we included two control non-painful situations in the list of 15 items in order to check if children are really able to imagine and rate the painfulness of hypothetical situations or systematically choose high scores whatever the situation. Fourth, we replaced the ‘smiling face’ with a ‘neutral face’ in the FS-3. Indeed, the low percentage utilization (see Table 2) of the lowest scores may result from the use of a smiling face as the ‘no pain’ anchor. Indeed, several studies have shown that children have significantly higher pain ratings when using scales that commence with smiling ‘no pain’ faces in comparison to FSs with neutral ‘no pain’ faces (Chambers and Craig, 1998; Chambers et al., 1999; Chambers et al., 2005). Children may be reluctant to point to a smiling ‘no pain’ face even if the situation is not painful probably because there is nothing happy about the circumstances even if pain is absent. This may have skewed the ratings towards faces at the upper end of the scale. With these four modifications, we commenced the second study.

3.1. Methods

3.1.1. Subjects

Seventy-six other children (41 boys and 35 girls) also aged from four to seven years participated in this second study. The recruitment procedure was identical to that of Study 1. The sample of children was similarly divided into two age categories: 4 to 5 years, n = 36 (22 boys, 14 girls) and 6 to 7 years, n = 40 (19 boys, 21 girls).

3.1.2. Faces scale

A FS-3 with a neutral face as ‘no pain’ anchor was presented to the children (Figure 1, (c)). This FS-3 was similar in face design to those used in Study 1.

3.1.3. Procedure

All participants were tested individually in the same conditions as Study 1. Again, before testing, a few minutes were spent in order to meet the child, make sure that he had correct notions of pain and to explain the task. Then, children were presented a series of 15 cartoon pictures portraying painful situations. For a list of the pictures, see Table 1, bottom list. Ten situations were selected from the items used in Study 1 on the basis of the analysis of responses and of children's understanding of the situations (items 1 to 10). Three painful situations were added to the list to extend the range of painfulness (items 11 to 13) and two non-painful situations were added as a control to check that children do not unthinkingly select high scores (items 14 and 15). Item 15 differs according to the sex of the child: to play soccer for boys and to skip a rope for girls. Vignettes were presented successively and in random order. Vignettes presented to girls showed a kid with long hair and a dress. Vignettes presented to boys showed a kid with short hair and shorts. All other elements of the vignette were perfectly similar. The facial expression of the depicted child was neutral. The child was asked to pretend that he was the child in the picture and to choose on the FS-3, the face that best reflects the amount of pain he would feel in each situation. The session lasted approximately 15 minutes. Test-retest stability of responses was investigated by resubmitting the test after an interim of ten days. Retest also served to further validate results found the first time. From the 76 children tested at the time of the first assessment, four could not be reassessed because of their absence from school at the time of the second visit.

3.1.4. Rasch analysis

The analysis of responses was performed exactly as for Study 1. Additionally, test-retest stability for item hierarchy and for children's pain representation was determined by an intraclass correlation coefficient.

3.2. Results

Separated Rasch analyses were conducted for each age group and for each assessment (test and retest). Items 14 and 15 elicited responses in the first response category from almost all children (74 out of 76 children at test time and 71 out of 72 children at retest time for item 14 and 62 out of 76 children at test time and 70 out of 72 children at retest time for item 15). These two control items were removed from the subsequent analyses because these items had been added to check that children did not unthinkingly select high scores and because items with the minimal score cannot be used for investigating the functioning of response scales. Distributions of child measures and item locations at test time are presented in Figure 3 (two right panels). Analyses led to item distributions that spread over 1.95 logit at test and over 2.36 logits at retest for the 4-5 years. For the group with 6-7 years of age item distributions spread over 4.33 logits at test and over 6.04 logits at retest. Compared with Study 1, the item locations are spread over a range about twice as wide for the 4-5 years old and five times as wide for the 6-7 years old. In other words, children's ability to differentiate the different situations according to their painfulness is considerably improved in this second study. The comparison of the two age groups highlights the wider range of item locations for the older children compared with the younger or in other words the higher capacity of the older ones for distinguishing the different situations according to their painfulness. On the other hand, children still favor high pain scores as revealed by the percentage utilization of each response category (Table 3). As for Study 1, this percentage increases together with the response categories and is very similar at test and at retest time. The test-retest stability of situation locations is presented in Figure 5 (upper panels). Situation locations are highly correlated between both assessments (intraclass coefficient correlation = 0.89, $p < 0.001$ for the 4-5 years and intraclass coefficient correlation = 0.95, $p < 0.001$ for the 6-7 years). On the other hand, children measures estimated by the model at the first and the second assessment are poorly but nevertheless significantly correlated (intraclass coefficient correlation = 0.47, $p = 0.0041$ for the 4-5 years and intraclass coefficient correlation = 0.44, $p = 0.0021$ for the 6-7 years) (Figure 5, bottom panels).

Table 3. Percentage utilization of categories and proportion of items with ordered thresholds.

Age (years)	*N*	% utilization of categories						Proportion of items with ordered thresholds	
		test			retest			test	retest
		0	1	2	0	1	2		
4-5	36	25	28	47	26	28	46	8/13	10/13
6-7	40	20	39	41	19	36	45	13/13	12/13

The examination of category functioning is summarized in Table 3. The proportion of items presenting ordered thresholds is much higher than for Study 1. For the 4-5 year-olds, this proportion is equal to 8/13 at test time and to 10/13 at retest time. For the 6-7 year-olds, this proportion is maximal at test time (13/13) and nearly maximal at retest time (12/13). The mean distance between both thresholds also significantly increases with age at test (paired *t*-test, $p < 0.001$) (Figure 6) and retest time (paired *t*-test, $p = 0.005$). The extremely narrow range for which category 1 is the most probable in the 4-5 year-olds signifies that this category is not really distinguished. These results confirm that children of 4-5 years old only distinguish two response levels. On the other hand, the 6-7 year olds clearly distinguish all three levels of the FS-3 even better than in Study 1.

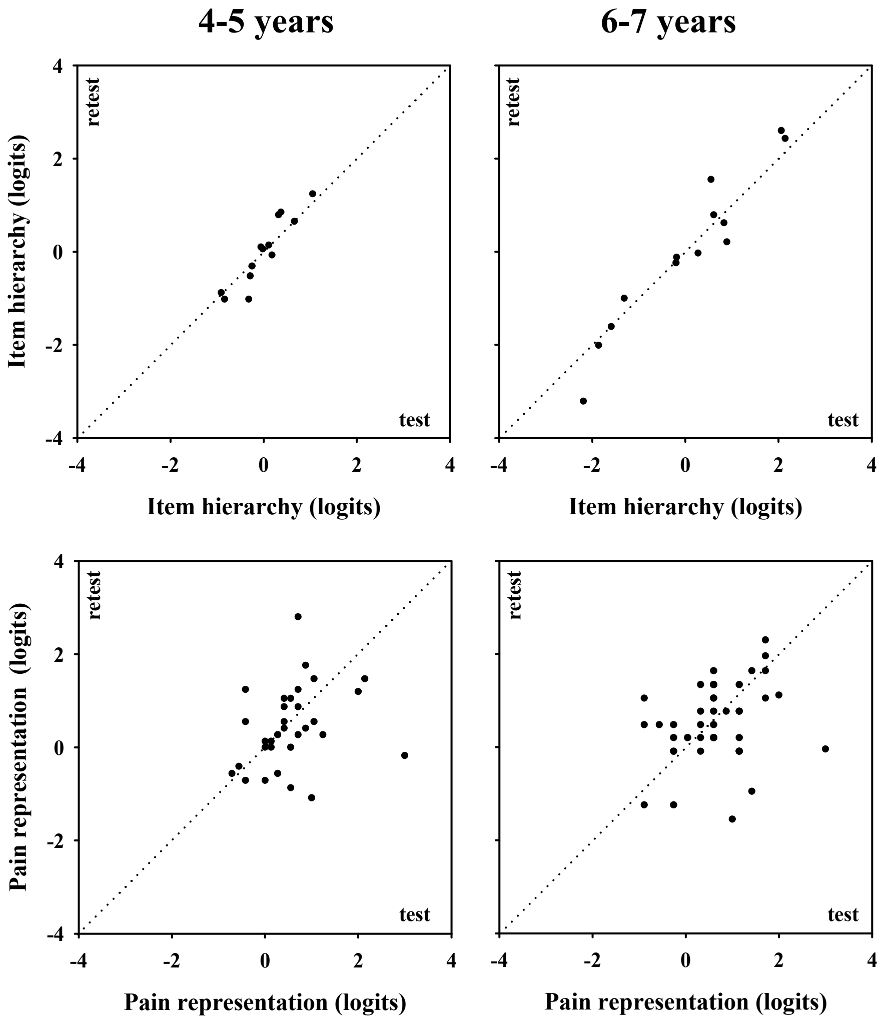


Figure 5. Upper panels: relationship between item painfulness across time for the 4-5 year-old (left panel) and for the 6-7 year-old children (right panel). Most painful items are plotted in the bottom/left part of the panels. Bottom panels: relationship between the pain representation measures of children at the first and second assessment (delay: 10 days) for the 4-5 year-old (left panel) and for the 6-7 year-old children (right panel). Children with a high representation of pain intensity are plotted in the top/right part of the panels.

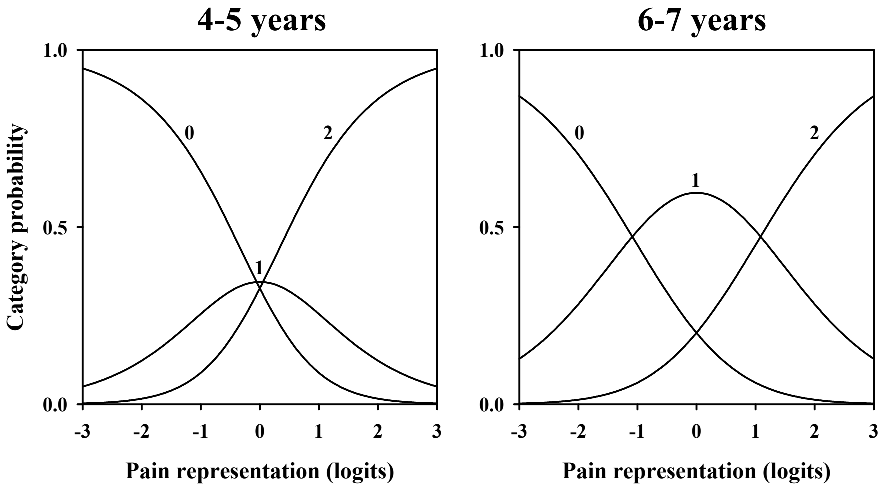


Figure 6. Category probability curves averaged across all items rated on the 3-level FS at test time for both age groups.

4. Discussion

The number of response categories on a pain intensity scale is increased with the intent of obtaining a more sensitive indication of pain severity. However, beyond a certain number of categories, confusion appears, as the different response categories are not distinguished anymore, compromising the validity of the ratings. The aim of the present studies was to determine the number of faces constituting the best trade off between precision and discrimination capacity in children. In the first study, children between four and seven years were asked to rate 25 painful situations on a 6- and a 3-level FS with a smiling face as the ‘no pain’ anchor. This first study revealed the low performance of the FS-6 as compared to the FS-3 and also the difficulty children experienced in discriminating the different situations according to their painfulness. Consequently, the second study was conducted with the objective of improving children’s performance in discriminating between situations. To improve the representation of the situations, pictures accompanied them. To keep attention focused, the number of situations was reduced to 15. To enhance the use of the first category of the FS-3, a neutral face replaced a smiling face as the ‘no pain’ anchor. In addition, participants were tested twice to examine the stability of responses.

Study 2 led to a great improvement in children's capacity to differentiate the different situations according to their painfulness. This improvement was greater for the older age group indicating that the ability to differentiate painful situations not only depend on the way they are presented but mostly on the age of the child. The use of a faces scale with a neutral face as the 'no pain' anchor in Study 2 only slightly raised the response rate in the lower intensity categories. The percentage utilization still increased with the response categories. Other factors must be invoked to explain the lack of low ratings. Probably, the highest contributing factor is the small number of barely painful situations. Indeed, the two least painful situations (item 4 and 12) produced a pattern of score frequencies that decreases with the response categories indicating that the addition of such situations would have increased the percentage utilization of the lowest scores. Another factor explaining the lack of low scores is the previously reported age-related difference in children's pain report. Several studies have shown that young children (about the same age as those enrolled in the present research) report more pain than older children (Arts et al., 1994; Goodenough et al., 1997; Chambers and Craig, 1998). However, the reasons for these higher scores are not clear. The cause may be a rating bias due to cognitively immature judgment of pain intensity and/or of usage of rating scales. The bias may also be due to true age difference in children's perceptions of painful events due to the activation of more nociceptors (spatial and temporal summation) in smaller bodies as discussed in Arts and coworkers (1994). This rating bias of young children towards the upper end of the response scale could partly explain the reported parental underestimation of children's pain (Chambers et al., 1999; Chambers et al., 2005).

Ratings provided by children at the first and at the second assessment resulted in item locations that are highly correlated for each age group. This signifies that, when considering the responses of all children, items are consistently rated across time. The most painful items at test time are still estimated as the most painful at retest time and *vice versa*. On the other hand, the same ratings resulted in measures of child pain representations that are significantly but weakly correlated for the 4-5 year-olds as well as for the 6-7 year-olds. This signifies that each individual child rated the different situations moderately consistently from one time to another. Children probably have difficulty imagining such hypothetical events despite the help of pictures. More than adults, their pain representation is also likely to be influenced by external factors such as the emotional status and the level of concentration at the time of the assessment.

In Study 2, modifications in the protocol resulted in an improvement of the discrimination of the response categories of the FS-3 when assessing the pain intensity of hypothetical painful events. In the 4-5 year-old group, a majority of items (8/13 and 10/13 at test and retest time respectively) presented ordered thresholds. Nevertheless, the mean distance between thresholds is extremely small (0.04 logits at test time and 0.33 logits at retest time). For the 6-7 year-old children, at test time and at retest time, almost all items displayed ordered thresholds (13/13 and 12/13, respectively). Furthermore, for this age group, the mean distance between thresholds is much wider (2.17 logits at test time and 2.02 logits at retest time) indicating that the middle category is well distinguished. Linacre (2002) recommended a minimal distance of 1.4 logit in order for a 3-category scale to be equivalent to two dichotomies. On the whole, these results show that 4-5 year-old children distinguish only two response categories when rating imaginary painful situations on a faces scale and that 6-7 year-old children are able to use the FS-3 for the same task. The difficulty of young children to use FSs is consistent with findings of recent research (Stanford et al., 2006). This study examined young children's ability to use the 6-level Faces Pain Scale-Revised (Hicks et al., 2001) in response to vignettes and investigated the role of developmental factors (numerical reasoning, language and overall cognitive development) in predicting their ability to use the scale. Authors calculated an error score for each child based on unexpected responses of pain severity depicted in the vignettes. They found that many young children aged 3-6 years experienced difficulties using the Faces Pain Scale-Revised and that age was the only significant predictor of children's ability to use the scale. They explained this latter result by the fact that age encompasses several cognitive domains of child development. The present study suggests that this reported difficulty to use a FS is probably due to difficulty in discriminating the six different levels of responses and that reduction in the number of faces could improve the validity of children's responses.

Finally, several important directions for future research can be recommended. An important next step is refining the developmental progress in using FSs by using a larger sample of participants, which would allow finer division of the sample according to age. Another point to investigate is the possible ability of 6 and 7 year-old children to distinguish four and five response levels on a face scale. An exploratory analysis performed on responses of the FS-6 used in Study 1 indicated that the 6-7 year-old children could distinguish four response categories. These analyses consisted in grouping some categories from the FS-6 and in reanalysing data. The grouping procedure where face 2 is grouped with face 3 and face 4 is grouped with face 5 led to satisfying results in terms of number of items presenting ordered thresholds (24/25) and in terms of mean distance between thresholds (1.04 logit). However, these *a posteriori* combinations are somewhat artificial in the sense that the investigator cannot be sure that children would use the new FS-4 as postulated by the combining of categories. Indeed, even if some studies have shown that the characteristics of the optimal categorization identified by the Rasch post-hoc analysis in a previous study (Zhu et al., 2001) could be maintained when the revised scale was applied to the same populations (Zhu, 2002), each new combination must be validated empirically.

To conclude, appropriate management of pain in children depends on valid, reliable and age-appropriate pain measures. In the past two decades, there has been an increased interest and significant improvements in pediatric pain measurement with, among others, the development of FSs. However, the central issue of the discrimination of the response categories on these scales has not been thoroughly studied. The present studies were conducted to answer this question. From their results, it appears that children from four to seven do not distinguish as many faces as proposed by the majority of available FSs when rating hypothetical situations. If these results are confirmed in a clinical pain context, they will have important implications. For instance, since young children (before six) appear dichotomous, the more valid self-assessment of pain is maybe simply to ask them if they are in pain. Another solution would be to use parents' perception of children's pain as we can suppose that it will be more accurate.

Conclusion and perspectives

In the present work, the Rasch methodology was successfully used to solve practical pain measurement problems encountered with self-report rating scales: the transformation of ordinal scores into interval-level measures, the development and validation of a new measurement tool and the investigation of rating scale functioning. Rasch analysis was chosen because it provides fundamental measurement, the type of measurement taken for granted in the physical sciences (Bond and Fox, 2001; Hobart, 2002; Massof, 2002; Andrich, 2004).

In the first chapter, the basics of the methodology have been presented and applied to the measurement of experimentally induced pain. Several intensities of CO_2 laser heat stimulation were delivered on the hand of each subject who had to rate their pain perception on a verbal rating scale. This experimental design was chosen because it is typical of protocols generally used to examine the effectiveness of a therapeutic intervention. This first chapter stressed the ability of the Rasch models to transform the original ordinal scores into linear measures on an interval scale. Linearity of the measurement scale is not obtained with ordinal scores. First, because ordinal rating scales have potentially unequal intervals between response categories. Second, because total scores rest on the assumption that all items are parallel instruments, with the same weight in the summation procedure. And yet, the same score obtained on different items does not necessarily represent the same amount of the latent variable (Merbitz et al., 1989). For example, answering 'slightly painful' to a 400 mJ and to the 775 mJ stimulus are both scored as 1. However, both responses do not correspond to the same level of pain perception since both stimuli are not of the same intensity. The logistic transformation of ordinal total scores into interval-level measures with the Rasch model allows arithmetic operations (e.g., summation) and parametric statistics. But, implications of this transformation go far beyond mathematical and statistical aspects. Most significantly, the interval-level properties of pain perception measures have relevance in clinical practice. They imply that a difference of one unit of measurement between two patients represents the same amount of pain throughout the scale allowing meaningful quantitative comparisons between and within patients. This quantification provides rigorousness to clinical trials evaluating pain treatment.

In chapter 2, a second application of the Rasch methodology was presented: the development and validation of a new measurement tool. In that chapter, an item questionnaire, the Situational Pain Scale (SPS), measuring the mental representation of pain in imaginary painful situations was developed and validated with healthy and chronic pain adults. The Rasch model was successfully used to select 18 items constituting a unidimensional and invariant measurement scale. In chronic pain patients, the pain representation measures were significantly related to present pain intensity and anxiety about pain. Fibromyalgia patients showed significantly higher pain representation measures than other groups of patients as well as healthy subjects. The pain representation as measured by the SPS was quite stable over time in both healthy subjects and chronic pain patients.

The next question to be investigated is that of the evolution of this pain representation from an acute to a chronic pain condition. We may hypothesize that the chronification of an acute pain will modify this representation but it remains to be empirically proven. Association between the SPS and social variables such as level of education and the household income should also be examined. In addition, future work is needed to study the value of the SPS in predicting patients behaviors. Several hypotheses that should be examined are proposed. First, the predictive value of the SPS in detecting among acute pain patients those more at risk of developing long-term problems. In this regard, a population of low back pain sufferers could be particularly appropriate. Second, the SPS usefulness in predicting future coping strategies developed by the chronic pain patient such as avoidance behaviors could be determined. Third, the ability of the SPS to detect among preoperative patients those who will require more care and possibly analgesia in the post-operative period could be investigated.

The discussion of chapter 2 advances the idea that the measure of the pain representation of imaginary situations obtained on the SPS largely reflects the attitude of the subject towards pain. This attitude is the result of a series of psychological variables influencing the perception of pain such as current emotional status, painful past experiences and socio-cultural contexts. Another questionnaire assessing this attitude, the Pain Attitude Questionnaire (PAQ) has been developed with healthy individuals (Yong et al., 2001) and revised in adult patients with chronic pain (Yong et al., 2003). Future studies should investigate the degree of association between the SPS and the PAQ. Nevertheless, we may already highlight a major difference in the philosophy behind the development of both instruments. In the PAQ, pain attitude is assessed by submitting rather direct statements (e.g., 'I think I can tolerate more pain than other people') to subjects who have to rate their level of agreement (or disagreement) with each of them. These kinds of statements suppose that subjects are totally conscious of their own attitude. The SPS, on the other hand utilizes an indirect method to assess the subjects' attitude towards pain via their rating of the pain intensity of imaginary situations. This procedure does not require that subjects are aware of their attitude towards pain.

In Rasch measurement, polytomous response format allows the choice of parameterization between rating scale and partial credit parameters. In the rating scale model, all items share the same rating scale structure. In the partial credit model, each item has its own rating scale structure. The partial credit model, by introducing more parameters into the estimation, generally provides a better fit of the data to the model. However, in both studies presented in chapter 1 and 2, a rating scale model was preferred to a partial credit model. Several reasons motivate this choice. First, in both studies, all items (laser stimuli in chapter 1 and imaginary painful situations in chapter 2) were presented to the respondents with the same response categories. Consequently, it is rather counter-intuitive to imagine that the response scale will change from one item to another. Changes in the rating scale structure from item to item make the scale more difficult to use and to comprehend for others. The second factor motivating the choice of the rating scale model is the higher estimate stability as compared to the partial credit model. The stability of the estimation depends on the number of observations in each category. In the rating scale model, as all items share the same structure of the response scale, the number of observations in each category is much greater and the estimation much more stable.

Chapter 3 detailed a third application of the Rasch models: the investigation and optimization of rating scale functioning. Two studies were conducted to answer the question of the optimal number of response levels required to assess the intensity of pain. A first one concerned verbal rating scales in a population of healthy adults while a second study investigated faces scales in a population of healthy children. Results showed that adults could distinguish around four levels of response while the youngest children (4-5 years old) distinguished two levels and older children (6-7 years old) appropriately used three levels. An interesting question that should be addressed by a future study is the possible ability of older children (6-7 years old) to distinguish four or five response levels.

Studies in chapter 3 involved a particular task: the assessment of imaginary painful situations. This testing situation was chosen as a first step to avoid submitting large samples of subjects (particularly children) with, for instance, an experimental nociceptive stimulus without empirical basis as regards the optimal number of response categories needed to assess the intensity of pain. Now that these studies have brought a precise response to this question, we may hypothesize that the optimal number of response levels will be relatively similar in other contexts and, in the future, relatively 'sure' protocols concerning the number of response categories to be proposed and investigated may be set up in clinical and experimental pain contexts.

The low ability to distinguish response levels found in these two studies is consistent with the function of the nociceptive system. From a biological perspective, pain is critical in promoting survival. The main function of the nociceptive system is to alert the organism to damaging or potentially damaging stimuli. In this regard, we may hypothesize that the human system of pain is characterized by a limited number of alarm levels allowing the human being to give the quickest and the most suitable answer to potentially damaging stimuli. The numbers of distinguishable response levels found in the present work are also consistent with those found for other attributes (Penta et al., 2001; Arnould et al., 2004; Brodersen, 2006) supporting the hypothesis of a general limit in the human capacity for discerning response levels on rating scales whatever the variable being assessed. For adults without major cognitive deficit, this number turns out to be around four with some variation according to the type of rating scale, the variable being rated and the subject population.

There is now substantial evidence, theoretical as well as empirical, to indicate the usefulness of Rasch models to solve diverse pain measurement problems. This work has presented several of its applications, namely the transformation of ordinal scores into interval-level measures, the development and validation of new measurement tools and the investigation of rating scales functioning. In addition, beyond the applications developed in the different chapters of the present work, Rasch models may be applied to many other measurement situations, for instance, item banking (Wright and Bell, 1984) or computer adaptive testing (Linacre, 2000). In the last 20 years, the number of papers that used Rasch analysis in health assessment has been growing. Applications in pain measurement are, however, still much more limited. Moreover, in the great majority of cases, these applications in pain measurement in particular and in health sciences in general are conducted in the contexts of research. The use of this methodology in clinical practice remains unusual and significant progress is needed if the aim is to spread the method in clinical pain contexts. The main reasons for the poor utilization of the Rasch methodology in clinical medicine are, in our opinion, the lack of information about the limitations and misuses of ordinal scores and, above all, the relative obscurity of the method compared with the handling of ordinal scores. A first step for solving the problem of the complexity of the method has been successfully achieved with the development and the increasing access to user-friendly software, such as Winsteps (Linacre and Wright, 1998) and RUMM (Andrich et al., 2003). In the years to come, the next step will consist in the development of easy to use routines that will automatically provide the patient's Rasch measure from the simple encoding of his raw responses to a previously Rasch-built instrument. An example of such routine already exists for the abilhand (Penta et al., 2001) and abilhand-kids (Arnould et al., 2004) questionnaires (http://www.abilhand.org). Routines for other scales in the rehabilitation domain including the SPS will soon be available (http://www.rehab-scales.org). They are essential to extend the use of Rasch-calibrated instruments in clinical practice.

Glossary

Ability. Location or level of successful performance of the person on the variable.

Average measure. For a particular category, average of the difference between the subject's and the item location across all observed responses in that category.

Bias. Cf. Differential item functioning.

Calibration. Procedure that estimates item and person locations on the measurement scale from the observed responses.

Category frequency. Percentage of occurrence of the category.

Coherence. Degree of consistency between observed and expected scores.

Confidence interval (CI). The 95% confidence interval is the range along the variable comprised between two standard errors below and two standard errors above the estimate of a given item or person. The true item or person measure has 95% of chance to effectively lie with the CI.

Dichotomous. Response format that has only two categories such as 'not painful'/'painful'.

Differential item functioning (DIF). Lack of invariance of item estimates across a factor characterizing the sample (e.g., sex) or across testing occasions.

Difficulty. Item parameter characterizing the location of the item characteristic curve along the variable.

Discrimination. Item parameter characterizing the slope of the item characteristic curve at the inflexion point.

Estimation. Process used to obtain the values of item and person parameters (or locations) from the matrix of reported responses.

Expected response or expected score. Predicted response according to the estimated person and item parameters.

Fit. Degree of similarity between the observed responses and the responses expected by the model.

Fit statistics. Indexes determining the fit of the data to the model.

Fundamental measurement. Fundamental measurement requires an ordering system and a process of concatenation, so that measurement units are equal.

Infit statistics. Information-weighted fit statistics. Fit statistics that are more sensitive to unexpected responses from persons with a measure close to the item measure.

Interval scale. Scale in which the difference between two graduations, the value of the unit of measurement, is known and constant all along the measurement scale. The 0 point is conventionally fixed.

Invariance. Maintenance of item or person estimates across different subgroups of persons or items respectively or across testing occasions.

Item characteristic curves (ICC). Ogival relationship relating the probability of success to an item to the latent trait.

Item Response Theory (IRT). Psychometric methods assessing latent traits that overcome limitations of traditional analysis performed on ordinal scales. IRT methods model the probability of an individual's response to an item by a function called item characteristic curve (ICC) that takes a S-shape.

Latent variable or latent trait. Characteristic or attribute that cannot be directly observed, hidden within the person. A latent variable can be inferred from the behaviors of the person.

Logit. The unit of measurement resulting from the Rasch transformation of ordinal scores into log odds ratios on an interval scale. One logit unit actually expresses a difference between two persons, two items or one person and one item. It corresponds, for example, to an increase in person location that increases the odds of success (a response scored 1 in a dichotomy 0/1) by a factor e^1 (2.71).

Many-facet model. Expansion of Rasch's original (two-facets) model in which supplementary factors, in addition to item difficulty and subject ability, systematically influencing the observed response, are estimated.

Matrix. Table with responses where rows (or columns) correspond to persons and columns (or rows) correspond to items.

Mean square. One way to express fit statistics. Mean square is a transformation of the residuals, the difference between the expected and the observed scores. Mean square has an expected value of 1.

Observed score or observed response. Reported response or score in its original state.

Ordinal scale. Scale in which observations are rank-ordered. An increasing score (e.g., 1, 2, 3) is allocated to these observations. The size of the distance between ranks is not necessarily equal.

Outfit statistics. Outlier-sensitive fit statistics. Fit statistics that are more sensitive to unexpected responses from persons with a measure far from the item measure.

Partial credit model. Expansion of Rasch's original dichotomous model used for the analysis of data scored with a polytomous response format. In the partial credit model, items may have different response categories and a set of individual threshold estimates is provided for each item.

Polytomous. Response format that has more than two categories.

Probabilistic. Characteristic of a model for which the outcome is expressed mathematically as a probability as all the influences on the outcome cannot be known.

Pseudo-chance level. Item parameter characterizing the lower asymptote of the item characteristic curve.

Rasch model. One-parameter IRT model. The only parameter describing the item is its difficulty.

Rating scale. Response format presenting categories that are ordered according to the level of the variable they define.

Rating scale model. Expansion of Rasch's original dichotomous model used for the analysis of data scored with a polytomous response format. In the rating scale model, all items have the same response categories and a unique set of threshold estimates applies to all items.

Raw score. Cf. observed score or observed response.

Separation Reliability index. Ratio between the true measure variance and the observed (true + error) measure variance. Proportion of observed variance that is not due to measurement error. It is bounded by 0 and 1. Person reliability index estimates the replicability of person location if another set of items measuring the same variable were given to that sample. Item reliability index estimates the replicability of item hierarchy if another sample of comparable ability answered the same items.

Specific objectivity. Property of a measurement scale implying that comparisons between individuals are independent of which particular items within the class considered have been used and, conversely, that comparisons between items are independent of which particular sample answered them.

Standard error (SE). Value associated to each person measure and each item location representing the precision of the estimation. It represents the range within which the true measure is expected to lie.

Stochastic. Cf. probabilistic.

Targeting. Matching between item and person locations.

Total score. For a given subject, the total score is the sum of his scores to all items. For a given item, the total score is the sum of scores of all subjects to that item.

Threshold. Location along the variable corresponding to an equal probability of endorsing two adjacent categories.

Two-parameter item response model. Item response model for which the parameters describing the item are its difficulty and its discrimination.

Three-parameter item response model. Item response model for which the parameters describing the item are its difficulty, its discrimination and its *pseudo-chance-level*.

Unidimensionality. Property of a measurement instrument that measures one single attribute at a time.

.

References

Andrich D. Relationships between the Thurstone and Rasch approaches to item scaling. Appl Psychol Meas 1978a;2:449-460.

Andrich D. A rating formulation for ordered response categories. Psychometrika 1978b;43:561-573.

Andrich D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Appl Psychol Meas 1978c;2:581-594.

Andrich D. Rasch models for measurement. Newbury Park, Ca: SAGE Publications, Inc; 1988.

Andrich D. Distinctions between assumptions and requirements in measurement in the social sciences. In: Keats JA, Taft R, Heath RA, Lovibond SH, editors. Mathematical and theoretical systems. North-Holland: Elsevier Science Publishers; 1989. p. 7-16.

Andrich D. Category ordering and their utility. Rasch Meas Trans 1996;9:464-465.

Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms. Med Care 2004;42:I7-I-16.

Andrich D, de Jong JHAL, Sheridan BE. Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost J, Langeheine R, editors. Applications of latent trait and latent class models in the social sciences. Münster: Waxmann; 1997. p. 59-72.

Andrich D, Sheridan B, Luo G. RUMM2020: Rasch unidimensional measurement models for measurement. Perth: RUMM Laboratory; 2003.

Arnould C, Penta M, Renders A, Thonnard J-L. ABILHAND-Kids: a measure of manual ability in children with cerebral palsy. Neurology 2004;63:1045-1052.

Arts SE, Abu-Saad HH, Champion GD, Crawford MR, Fisher RJ, Juniper KH, Ziegler JB. Age-related response to lidocaine-prilocaine (EMLA) emulsion and effect of music distraction on the pain of intravenous cannulation. Pediatrics 1994;93:797-801.

Bending AW. Reliability and the number of rating scale categories. J Appl Psychol 1954a;38:38-40.

Bending AW. Reliability of short rating scales and the heterogeneity of the rated stimuli. J Appl Psychol 1954b;38:38-40.

Bertrand R, Blais J-G. Modèles de mesure: l'apport de la théorie des réponses aux items. Sainte-Foy: Presses de l'Université du Québecq; 2004.

Beyer JE. The Oucher: a user's manual and technical report. Evanston, IL: The Hospital Play Equipment Company; 1984.

Bieri D, Reeve RA, Champion GD, Addicoat L, Ziegler JB. The Faces Pain Scale for the self-assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for ratio scale properties. Pain 1990;41:139-150.

Bjelland, I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res 2002;52:69-77.

Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah: Lawrence Erlbaum Associates; 2001.

Breivik EK, Björnsson GA, Skovlund E. A comparison of pain rating scales by sampling from clinical trial data. Clin J Pain 2000;16:22-28.

Brodersen J. Measuring psychosocial consequences of false-positive screening results – breast cancer as an example. Thesis dissertation, University of Copenhagen; 2006. p. 80-94.

Chambers CT, Craig KD. An intrusive impact of anchors in children's faces pain scales. Pain 1998;78:27-37.

Chambers CT, Giesbrecht K, Craig KD, Bennett SM, Huntsman E. A comparison of faces scales for the measurement of pediatric pain: children's and parents' ratings. Pain 1999;83:25-35.

Chambers CT, Hardial J, Craig KD, Court C, Montgomery C. Faces scales for the measurement of postoperative pain intensity in children following minor surgery. Clin J Pain 2005;21:277-285.

Chapman CR. Pain, perception, and illusion. In: Sternbach RA, editor. The psychology of pain. New York: Raven Press; 1986. p. 153-179.

Chapman CR, Casey KL, Dubner R, Foley KM, Gracely RH, Reading AE. Pain measurement: an overview. Pain 1985;22:1-31.

Choinière M, Melzack R, Girard N, Rondeau J, Paquin M-J. Comparisons between patients' and nurses' assessment of pain and medication efficacy in severe burn injuries. Pain 1990;40:143-152.

Clark WC. Pain sensitivity and the report of pain: an introduction to sensory decision theory. Anesthesiology 1974;40:272-287.

Conrad KJ, Smith EV. International conference on objective measurement: applications of Rasch analysis in health care. Med Care 2004;42:I-1-I-6.

Coulthard P, Pleuvry BJ, Dobson M, Price M. Behavioural measurement of postoperative pain after oral surgery. Br J Oral Maxillofac Surg 2000;38:127-131.

Crombez G, Eccleston C, Van den Broeck A, Goubert L, Van Houdenhove B. Hypervigilance to pain in fibromyalgia: the mediating role of pain intensity and catastrophic thinking about pain. Clin J Pain 2004;20:98-102.

Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, Williams JI. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol 2003;56:1076-1083.

Eich E, Reeves JL, Jaeger B, Graff-Radford SB. Memory for pain: relation between past and present pain intensity. Pain 1985;23:375-380.

Feine JS, Lavigne GJ, Thuan Dao TT, Morin C, Lund JP. Memories of chronic pain and perceptions of relief. Pain 1998;77:137-141.

Finn RH. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educ Psychol Meas 1972;32:255-265.

Fisher WP. Reliability statistics. Rasch Meas Trans 1992;6:238.

Fisher WP. The Rasch debate: validity and revolution in educational measurement. I In: Wilson M, editor. Objective measurement: theory into practice, volume 2. Norwood: Ablex; 1994. p. 36-72.

Hambleton RK, Swaminathan H, Rogers H J. Fundamentals of Item Response Theory. Newbury Park, Ca: Sage Publications, Inc; 1991.

Gaffney A, Dunne EA. Developmental aspects of children's definitions of pain. Pain 1986;26:105-117.

Gagliese L, Melzack R. Age-related differences in the qualities but not the intensity of chronic pain. Pain 2003;104:597-608.

Gaskin ME, Greene AF, Robinson ME, Geisser ME. Negative affect and the experience of chronic pain. J Psychosom Res 1992;36:707-713.

Gelman R, Baillargeon R. A review of some Piagetian concepts. In: Flavell JH, Markman EM, editors. Handbook of child psychology, volume 3: Cognitive development. New York: Wiley Press; 1983. p. 167-230.

Gibson SJ, Littlejohn GO, Gorman MM, Helme RD, Granges G. Altered heat pain thresholds and cerebral event-related potentials following painful CO_2 laser stimulation in subjects with fibromyalgia syndrome. Pain 1994;58:185-193.

Goodenough B, Kampel L, Champion GD, Laubreaux L, Nicholas MK, Ziegler JB, McInerney. An investigation of the placebo effect and age-related factors in the report of needle pain from venipuncture in children. Pain 1997;72:383-391.

Gracely RH, Dubner R. Pain assessment in humans - A reply to Hall. Pain 1981;11:109-120.

Gracely RH, McGrath P, Dubner R. Ratio scales of sensory and affective verbal pain descriptors. Pain 1978;5:5-18.

Granot M, Buskila D, Granovsky Y, Sprecher E, Neumann L, Yarnitsky D. Simultaneous recording of late and ultra-late pain evoked potentials in fibromyalgia. Clin Neurophysiol 2001;112:1881-1887.

Greenough CG, Fraser RD. Comparison of eight psychometric instruments in unselected patients with back pain. Spine 1991;16:1068-1074.

Guru V, Dubinsky I. The patient vs. caregiver perception of acute pain in the emergency department. J Emerg Med 2000;18:7-12.

Guttman L. The basis for scalogram analysis. In Stouffer SA, Guttman L, Suchman EA, Lazarsfield PF, Star SA, Clausen JA, editors. Measurement and prediction. Princeton: Princeton University Press; 1950. p. 60-90.

Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory, volume 2. Newbury Park: Sage Publications; 1991.

Hardy JD, Wolff HG, Goodell H. Pain sensations and reactions. Baltimore: Williams and Wilkins; 1952. p. 146-172.

Harke H, Gretenkort P, Ladleif HU, Rahman S. Spinal cord stimulation in sympathetically maintained complex regional pain syndrome type I with severe disability. A prospective clinical study. Eur J Pain 2005;9:363-373.

Herrmann C. International experiences with the Hospital Anxiety and Depression Scale. A review of validation data and clinical results. J Psychosom Res 1997;42:17-41.

Hicks CL, von Baeyer CL, Spafford PA, van Korlaar I, Goodenough B. The faces pain scale-revised: toward a common metric in pediatric pain measurement. Pain 2001;93:173-183.

Hobart J. Measuring disease impact in disabling neurological conditions: are patients' perspectives and scientific rigor compatible? Curr Opin Neurol 2002;15:721-724.

Hunter M, McDowell L, Hennessy R, Cassey J. An evaluation of the Faces Pain Scale with young children. J Pain Symptom Manage 2000;20:122-129.

Huskisson EC. Measurement of pain. The Lancet 1974;9:1127-1131.

International Association for the Study of Pain, Taxonomy Committee. Pain terms: a list with definitions and notes on usage. Pain 1979;6:249-252.

Jenkins GD, Taber TD. A Monte Carlo study of factors affecting three indices of composite scale reliability. J Appl Psychol 1977;62:392-398.

Jensen MP, Karoly P. Self-report scales and procedures for assessing pain in adults. In: Turk DC, Melzack R, editors. Handbook of pain assessment. New York: The Guilford Press; 1992. p. 135-151.

Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain 1986;27:117-126.

Jensen MP, Karoly P, O'Riordan EF, Bland F Jr, Burns RS. The subjective experience of acute pain: an assessment of the utility of 10 indices. Clin J Pain 1989;5:153-159.

Jensen MP, Turner JA, Romano JM. What is the maximum number of levels needed in pain intensity measurement? Pain 1994;58:387-392.

Keefe FJ, Williams DA, Smith SJ. Assessment of pain behavior. In: Turk DC, Melzack R, editors. Handbook of pain assessment. New York: The Guilford Press; 1992. p. 275-292.

Komorita SS, Graham WK. Number of scale points and the reliability of scales. Educ Psychol Meas 1965;25:987-995.

Kosek E, Ekholm J, Hansson P. Sensory dysfunction in fibromyalgia patients with implications for pathogenic mechanisms. Pain 1996;68:375-383.

Kremer E, Atkinson JH, Ignelzi RJ. Measurement of pain: patient preference does not confound pain measurement. Pain 1981;10:241-248.

Kuttner L, LePage T. Faces scales for the assessment of pediatric pain: a critical review. Can J Behav Sci 1989;21:198-203.

Laveault D, Grégoire J. Introduction aux théories des tests en sciences humaines. Paris: De Boeck Université; 2000.

LeBaron S, Zeltzer L. Assessment of acute pain and anxiety in children and adolescents by self-reports, observer reports, and a behaviour checklist. J Consult Clin Psychol 1984;52:729-738.

Ledingham J, Doherty S, Doherty M. Primary fibromyalgia syndrome - an outcome study. Br J Rheumatol 1993;32:139-142.

Linacre, JM. Many-facet Rasch measurement. Chicago: Mesa Press; 1989.

Linacre, JM. A user's guide to Facets: Rasch measurement computer program. Chicago: Mesa Press; 1994a.

Linacre, JM. Sample size and item calibration stability. Rasch Meas Trans 1994b;7:328.

Linacre JM. Categorical misfit statistics. Rasch Meas Trans 1995;9:450-451.

Linacre JM. Judging plans and facets (research note 3). Chicago: Mesa Press; 1997. Available: http://www.rasch.org/rn3.htm.

Linacre JM. Category disordering vs. step disordering. Rasch Meas Trans 1999;13:675-678.

Linacre JM. Computer-adaptive resting: a method whose time has come. Chicago: Mesa Press; 2000.

Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002;3:85-106.

Linacre JM, Wright BD. A user's guide to Winsteps: Rasch-model computer program. Chicago: Mesa Press, 1998.

Lissitz RW, Green SB. Effect of the number of scale points on reliability: a Monte Carlo approach. J Appl Psychol 1975;60:10-13.

Littman GS, Walker BR, Schneider BE. Reassessment of verbal and visual analog ratings in analgesic studies. Clin Pharmacol Ther 1985;38:16-23.

Lorenz J, Grasedyck K, Bromm B. Middle and long latency somatosensory evoked potentials after painful laser stimulation in patients with fibromyalgia syndrome. Electroencephalogr clin Neurophysiol 1996;100:165-168.

Massof RW. The measurement of vision disability. Optom Vis Sci 2002;79:516-552.

Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149-174.

Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert scale items? Study I: reliability and validity. Educ Psychol Meas 1971;31:657-674.

McArthur DL, Casey KL, Morrow TJ, Cohen MJ, Schandler SL. Partial-Credit modeling and response surface modeling of biobehavioral data. In: Wilson M, editor. Objective measurement: theory into practice (volume 1). Norwood, NJ: ABLEX Publishing Corporation; 1992. p. 109-120.

McArthur DL, Cohen MJ, Schandler SL. A philosophy for measurement of pain. In: Chapman CR, Loeser JD, editors. Issues in pain measurement. New York: Raven Press; 1989. p. 37-49.

McArthur DL, Cohen MJ, Schandler SL. Rasch analysis of functional assessment scales: an example using pain behaviors. Arch Phys Med Rehabil 1991;72:296-304.

McCracken LM, Zayfert C, Gross RT. The pain anxiety symptoms scale: development and validation of a scale to measure fear of pain. Pain 1992;50:67-73.

McDermid AJ, Rollman GB, McCain GA. Generalized hypervigilance in fibromyalgia: evidence of perceptual amplification. Pain 1996;112:1881-1887.

McGrath PA, Seifert CE, Speechley KN, Booth JC, Stitt L, Gibson MC. A new analogue scale for assessing children's pain: an initial validation study. Pain 1996;64:435-443.

McHorney CA, Monahan PO. Applications of Rasch analysis in health care. Med Care 2004;42:I-73-I-78.

Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain 1975;1:277-299.

Melzack R, Casey KL. Sensory, motivational, and central control determinants of pain: a new conceptual model. In: Kenshalo D, editor. The skin senses. Springfield: Chas C. Thomas. p. 423-439.

Melzack R, Katz J. Pain assessment in adult patients. In: Wall PD, Melzack R. Textbook of pain (5th edition). McMahon SB, Koltzenburg M, editors. Edinburgh: Churchill Livingstone; 2006. p. 291-304.

Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989;70:308-312.

Merskey H, Bogduk N. Classification of chronic pain, definitions of chronic pain syndromes and definitions of pain terms (2nd edition). Seattle: IASP Press, 1994.

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

Penta M, Arnould C, Decruynaere C, Thonnard J-L, Plaghki L. Développer et interpreter une échelle de mesure: applications du modèle de Rasch. Sprimont: Mardaga; 2005.

Penta M, Tesio L, Arnould C, Zancan A, Thonnard J-L. The ABILHAND questionnaire as a measure of manual ability in chronic stroke patients: Rasch-based validation and relationship to upper limb impairment. Stroke 2001;32:1627-1634.

Pesudovs K, Noble BA. Improving subjective scaling of pain using Rasch analysis. J Pain 2005;6:630-636.

Plaghki L, Delisle D, Godfraind JM. Heterotopic nociceptive conditioning and mental task modulate differently the perception and physiological correlates of short CO_2 laser stimuli. Pain 1994;57:181-192.

Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45-56.

Ramsay JO. The effect of number of categories in rating scales on precision of estimation of scale values. Psychometrika 1973;38:513-532.

Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960. (Expanded edition, Chicago: Mesa Press; 1980).

Sedlak K. Low-back pain. Perception and tolerance. Spine 1985;10:440-444.

Smith RM. Applications of Rasch Measurement. Chicago: Mesa Press, 1992.

Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000;1:199-218.

Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. J Outcome Meas 1998;2:66-78.

Spielberger CD, Gorsuch RL, Lushene RE. The State-Trait Anxiety Inventory (self-evaluation questionnaire). Palo Alto, Ca: Consulting Psychologists Press; 1970.

Stanford EA, Chambers CT, Craig KD. The role of developmental factors in predicting young children's use of a self-report scale for pain. Pain 2006;120:16-23.

Stanley MA, Beck JG, Zebb BJ. Psychometric properties of four anxiety measures in older adults. Behav Res Ther 1996;34:827-838.

Sternbach RA. Linguistic problems in pain measurement. In: Chapman CR, Loeser JD, editors. Advances in pain research and therapy (volume 12): issues in pain measurement. New York: Raven Press; 1989. p. 63-68.

Stevens SS. On the theory of scales of measurement. Science 1946;103:677-680.

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;125:143-157.

Symonds PM. On the loss of reliability in ratings due to coarseness of the scale. J Exp Psychol 1924;7:456-461.

Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:105-115.

Tesio L. Measurement in clinical vs. biological medicine: the Rasch model as a bridge on a widening gap. J Appl Meas 2004;5:362-366.

Tesio L, Granger CV, Fiedler RC. A unidimensional pain/disability measure for low-back pain syndromes. Pain 1997;69:269-278.

Thomeé R, Grimby G, Wright BD, Linacre JM. Rasch analysis of visual analog scale measurements before and after treatment of patellofemoral pain syndrome in women. Scand J Rehabil Med 1995;27:145-151.

Thorndike RM. IRT and intelligence testing: past, present, and future. In: Embretson SE, Hershberger SL, editors. The new rules of measurement: what every psychologist and educator should know. Mahwah: Lawrence Erlbaum Associates; 1999. p. 17-35.

Townsend JT, Ashby FG. Measurement scales and statistics: the misconception misconceived. Psychol Bull 1984;96:394-401.

Turk DC. Assessment of pain: the elusiveness of latent constructs. In: Chapman CR, Loeser JD, editors. Advances in pain research and therapy (volume 12): issues in pain measurement. New York: Raven Press; 1989. p. 267-279.

Vierck C J. Mechanisms underlying development of spatially distributed chronic pain (fibromyalgia). Pain 2006;24:242-263.

Wallenstein SL, Heidrich G 3rd, Kaiko R, Houde RW. Clinical evaluation of mild analgesics: the measurement of clinical pain. Br J Clin Pharmac 1980;10:319S-327S.

White LJ, Craig PT, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil 2002;83:822-831.

White KP, Nielson WR, Harth M, Ostbye T, Speechley M. Chronic widespread musculoskeletal pain with or without fibromyalgia: psychological distress in a representative community adult sample. J Rheumatol 2002;29:588-594.

Wolfe F. Pain Extend and Diagnosis: Development and validation of the regional pain scale in 12,799 patients with rheumatic disease. J Rheumatol 2003;30:369-378.

Wong DL, Baker CM. Pain in children: a comparison of assessment scales. Pediatr Nurs 1988;14:9-17.

Wright BD. Reliability and separation. Rasch Meas Trans 1996;9:472.

Wright BD, Bell SR. Item banks: what, why, how. J Educ Measure 1984;21:331-345.

Wright BD, Linacre M. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70:857-860.

Wright B, Linacre JM. Combining and splitting categories. Rasch Meas Trans 1992;6:233-235.

Wright BD, Masters GN. Rating scale analysis. Chicago: Mesa Press, 1982.

Wright BD, Panchapakesan N. A procedure for sample-free item analysis. Educ Psychol Meas 1969;29:23-48.

Wright BD, Stone MH. Best test design. Chicago: Mesa Press; 1979.

Yang JC, Wagner JM, Clark WC. Psychological distress and mood in chronic pain and surgical patients: a sensory decision theory analysis. In: Bonica J, Lindblom U, Igg A, editors. Advances in Pain Research and Therapy. New York: Raven Press; 1983. p. 901-906.

Yong HH, Bell R, Workman B, Gibson SJ. Psychometric properties of the Pain Attitudes Questionnaire (revised) in adult patients with chronic pain. Pain 2003;104:673-681.

Yong HH, Gibson SJ, Horne DJ, Helme RD. Development of a pain attitudes questionnaire to assess stoicism and cautiousness for possible age differences. J Gerontol B Psychol Sci Soc Sci 2001;56:279-84.

Zambito A, Bianchini D, Gatti D, Viapiana O, Rossini M, Adami S. Interferential and horizontal therapies in chronic low back pain: a randomized, bouble blind, clinical study. Clin Exp Rheumatol 2006;24:534-539.

Zhu W. A confirmatory study of Rasch-based optimal categorization of a rating scale. J Appl Meas 2002;3:1-15.

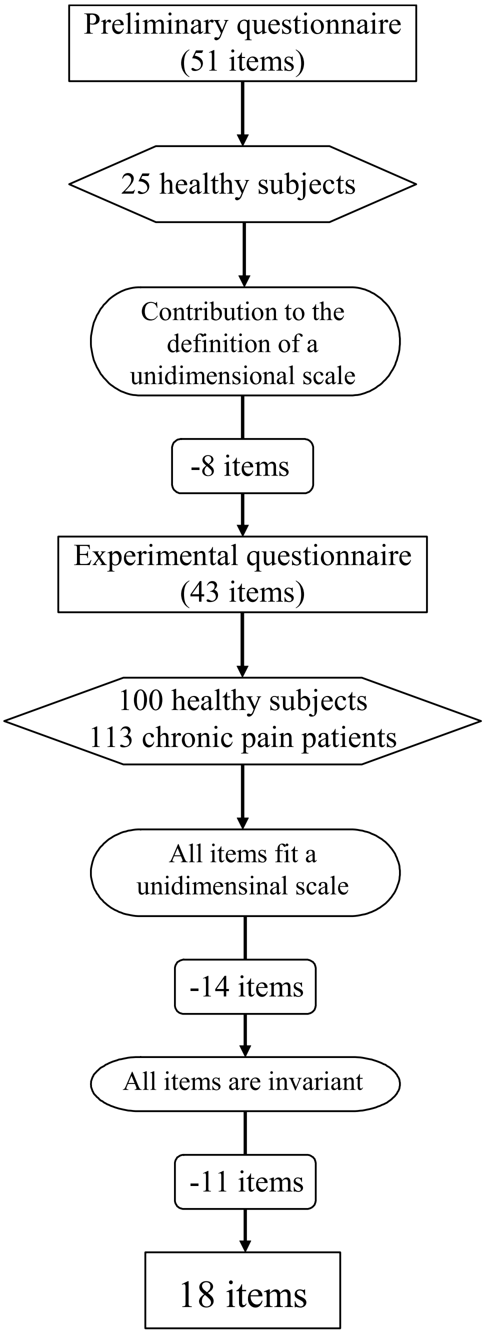
Zhu W, Timm G, Ainsworth B. Rasch calibration and optimal categorization of an instrument measuring women's exercise perseverance and barriers. Res Q Exerc Sport 2001;72:104-16.

Zhu W, Updyke WF, Lewandowski C. Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. J Outcome Meas 1997;1:286-304.

Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scan 1983;67:361-370.

Appendices

Appendix I.
Item reduction process of SPS (chapter 2)



Appendix II.
Preliminary version of SPS (chapter 2)

Items of the preliminary questionnaire*	Reason for item deletion
I get bitten by a mosquito	
I pick up a hot pan without gloves	
I cut myself with a sheet of paper	
I get a tooth drilled without a painkiller	Misfit
I have stomach aches	
I get sunburned on my face	
I have a raging toothache	
I walk barefoot along a gravel path	
My sinuses are clogged up	Overfit
I disinfect a sore	
I lie in one position on a hard mattress for an hour	Overfit
I fall from two-meter stairs	Overfit
I accidentally hit my thumb with a hammer	
My throat is examined by the doctor	
I move my foot after it falls asleep	
I get stung by a bee	
I get a speck of dust in the eye	
My eyes weep peeling onions	
I get splashed with melted butter when cooking	
I twist my ankle slightly	
I burn my tongue tasting scorching hot food	
I walk on burning sand	
My lips are chapped	
I step barefoot on a piece of sharp glass	
My fingers are frozen	
I get shampoo in my eye	
I untangle my hair with a comb	Misfit
I hit my funny bone	
I wrench my back lifting a heavy weight	
I get an injection in the arm	
Someone pulls my hair	
I get my fingers caught in the car door	
I have a mouth ulcer	
I have a stitch in my side	

* The 43 items of the experimental version of SPS are in bold type.

Items of the preliminary questionnaire*	Reason for item deletion
My ears pop from a change in altitude in an airplane	Misfit
I knock my head on the corner of a piece of furniture	
I get indigestion after a large meal	
I catch my finger in a zipper	
I spill some salt on a cut	Overfit
I get sunburned and someone touches me on that spot	
My teeth are examined by a dentist in a check-up	
My shower suddenly turns ice-cold	
The doctor takes my blood pressure	
I cut my hand with a sharp knife	
I stub my toe on a chair leg	
I bite my tongue	
I have a splinter under the skin of one finger	
I hold a very heavy bag in my hand	
I get out of breath trying to catch a bus	Misfit
I have a cramp in my foot	
I spill some boiling water on my hands	

* The 43 items of the experimental version of SPS are in bold type.

Appendix III.
Experimental version of SPS (chapter 2)

Items of the experimental questionnaire*	Reason for item deletion
I get bitten by a mosquito	Lack of invariance
I pick up a hot pan without gloves	Misfit
I cut myself with a sheet of paper	
I have stomach aches	Lack of invariance
I get sunburned on my face	
I have a raging toothache	Misfit
I walk barefoot along a gravel path	Lack of invariance
I disinfect a sore	
I accidentally hit my thumb with a hammer	Misfit
My throat is examined by the doctor	Misfit
I move my foot after it falls asleep	Misfit
I get stung by a bee	Lack of invariance
I get a speck of dust in the eye	
My eyes weep peeling onions	Lack of invariance
I get splashed with melted butter when cooking	Lack of invariance
I twist my ankle slightly	Lack of invariance
I burn my tongue tasting scorching hot food	
I walk on burning sand	
My lips are chapped	
I step barefoot on a piece of sharp glass	Misfit
My fingers are frozen	Misfit
I get shampoo in my eye	
I hit my funny bone	
I wrench my back lifting a heavy weight	Lack of invariance
I get an injection in the arm	
Someone pulls my hair	
I get my fingers caught in the car door	
I have a mouth ulcer	Lack of invariance
I have a stitch in my side	Overfit
I knock my head on the corner of a piece of furniture	
I get indigestion after a large meal	Misfit
I catch my finger in a zipper	
I get sunburned and someone touches me on that spot	
My teeth are examined by a dentist in a check-up	Misfit

* The 18 items of the final version of SPS are in bold type.

Items of the experimental questionnaire*	Reason for item deletion
My shower suddenly turns ice-cold	Misfit
The doctor takes my blood pressure	Lack of invariance
I cut my hand with a sharp knife	Misfit
I stub my toe on a chair leg	
I bite my tongue	
I have a splinter under the skin of one finger	
I hold a very heavy bag in my hand	Misfit
I have a cramp in my foot	Lack of invariance
I spill some boiling water on my hands	Misfit

* The 18 items of the final version of SPS are in bold type.

Appendix IV.
English version of SPS (chapter 2, 3.1 and 3.2)

SPS-Pain representation measure
English version

Name:_________________

	Could you estimate the intensity of pain during the following situations?	Not painful	Slightly painful	Moderately painful	Extremely painful	?
1	I disinfect a sore					
2	I get an injection in the arm					
3	Someone pulls my hair					
4	I hit my funny bone					
5	I get a speck of dust in the eye					
6	I knock my head on the corner of a piece of furniture					
7	I cut myself with a sheet of paper					
8	I bite my tongue					
9	I stub my toe on a chair leg					
10	I get sunburned and someone touches me on that spot					
11	I catch my finger in a zipper					
12	I have a splinter under the skin of one finger					
13	I get shampoo in my eye					
14	I get sunburned on my face					
15	I burn my tongue tasting scorching hot food					
16	I get my fingers caught in the car door					
17	My lips are chapped					
18	I walk on burning sand					

Appendix V.
French version of SPS (chapter 2, 3.1 and 3.2)

SPS-Pain representation measure
French version

Nom:__________

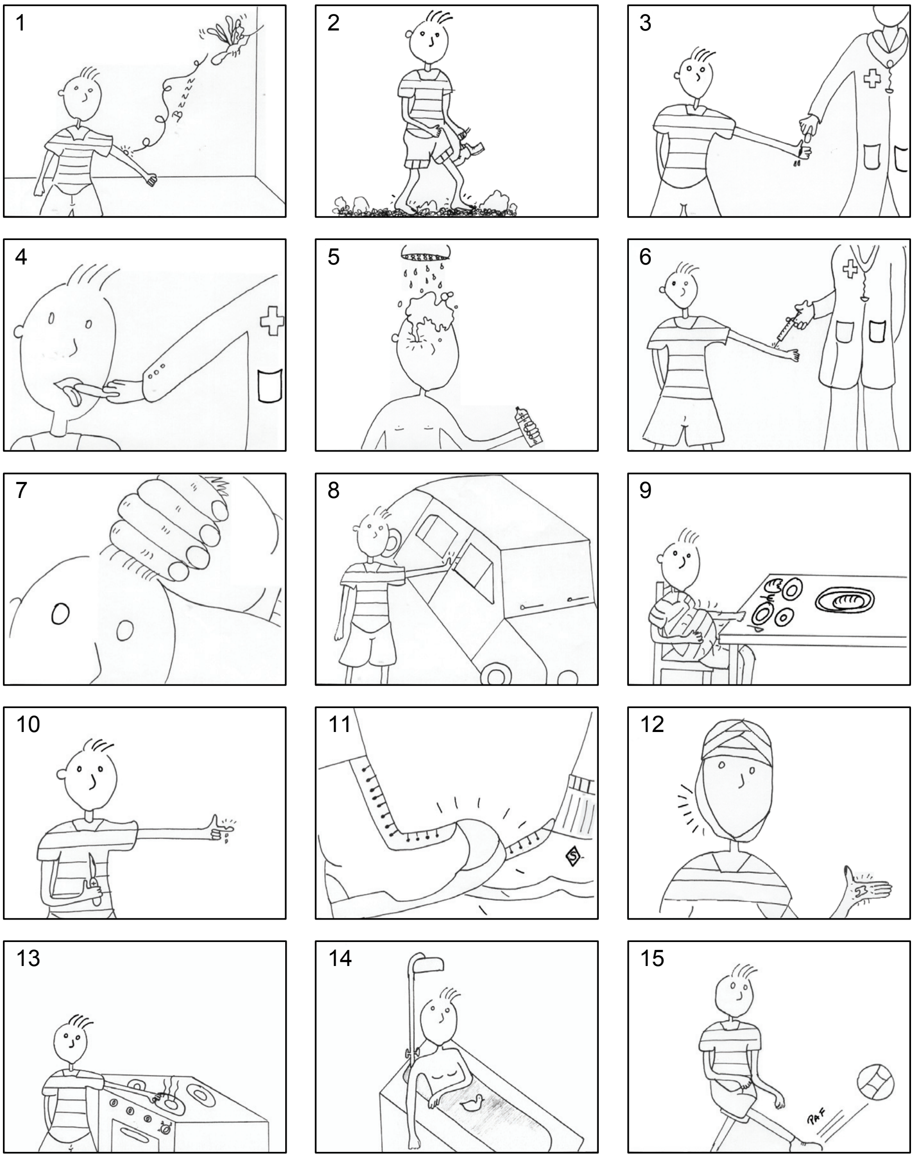
	Pouvez-vous estimer l'intensité de la douleur lors des situations suivantes?	**Pas douloureux**	**Légèrement douloureux**	**Moyennement douloureux**	**Extrêmement douloureux**	**?**
1	Je désinfecte une plaie					
2	Je reçois une injection dans le bras					
3	Quelqu'un me tire les cheveux					
4	Je me cogne le coude et ressens une décharge dans la main					
5	J'attrape une poussière dans l'oeil					
6	Je me cogne la tête sur le coin d'un meuble					
7	Je me coupe avec une feuille de papier					
8	Je me mords la langue					
9	Je me cogne l'orteil sur le coin d'une chaise					
10	J'ai un coup de soleil et quelqu'un me touche à cet endroit					
11	Je me pince le doigt dans une fermeture éclair					
12	J'ai une écharde sous la peau d'un doigt					
13	J'ai du shampoing dans l'œil					
14	J'ai un coup de soleil sur le visage					
15	Je me brûle la langue en goûtant un aliment trop chaud					
16	Je me coince les doigts dans une portière de voiture					
17	J'ai les lèvres gercées					
18	Je marche sur du sable brûlant					

Appendix VI.
English and French versions of the 25 painful situations used in chapter 3.3 (study 1)

English version	French version
A mosquito bit	Une piqûre de moustique
Getting sunburned on the face	Un coup de soleil sur le visage
A toothache	Une rage de dent
Walking barefoot on pebbles	Marcher pieds nus sur des cailloux
Cleaning a scratch	Nettoyer un bobo
The doctor examine the throat	Le docteur examine la gorge
Being stung by a bee	Une piqûre de guêpe
Getting a speck of dust in the eye	Une poussière dans l'œil
Burning one's tongue tasting scorching hot food	Se brûler la langue en goûtant un aliment trop chaud
Walking on burning sand	Marcher sur du sable brûlant
Getting one's fingers frozen	Les doigts gelés par le froid
Getting shampoo in the eye	Avoir du shampoing dans l'œil
Getting an injection at the doctor's	Recevoir une piqûre chez le médecin
Someone pulls the hair	Quelqu'un tire les cheveux
Getting the fingers caught in the car door	Avoir les doigts coincés dans une portière de voiture
Knocking the head on the corner of a piece of furniture	Se cogner la tête sur le coin d'un meuble
Getting indigestion after a large meal	Avoir une indigestion après un repas copieux
Catching fingers in a zipper	Se pincer les doigts dans une tirette
Getting sunburned and being touched on that spot	Avoir un coup de soleil et quelqu'un te touche à cet endroit
The dentist examines the teeth in a check-up	Le dentiste examine les dents
Cutting the hand with a knife	Se couper la main avec un couteau
Stubbing one's toe on a chair leg	Se cogner l'orteil sur le coin d'une chaise
Biting one's tongue	Se mordre la langue
Carrying a heavy bag	Porter un sac lourd
Spilling some boiling water on the hands	Renverser de l'eau bouillante sur les mains

Appendix VII.
Vignettes situations used in chapter 3.3 (study 2)

Vignettes situations presented to boys:



Vignettes situations presented to girls: