Scaling properties of pain intensity ratings in paediatric populations using the Faces Pain Scale-revised: Secondary analyses of published data based on the item response theory.

AVIAN Alexander[1,2,3]

MESSERER Brigitte[4]

FREY Andreas[3,5]

MEISSNER Winfried[2]

WEINBERG Annelie[6]

RAVEKES William[7]

BERGHOLD Andrea[1]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria

alexander.avian@medunigraz.at

andrea.berghold@medunigraz.at

[2]Department of Anesthesiology and Intensive Care, Jena University Hospital, Jena – Lobeda, Germany

Winfried.meissner@med.uni-jena.de

[3]Department of Research Methods in Education, Friedrich Schiller University Jena, Am Planetarium 4, 07743 Jena, Germany

andreas.frey@uni-jena.de

Alexander.avian@uni-jena.de

[4]Division of Anesthesiology for Cardiovascular Surgery and Intensive Care Medicine, Medical University of Graz, Auenbruggerplatz 29, 8036 Graz, Austria

brigitte.messerer@medunigraz.at

[5]Centre for Educational Measurement (CEMO) at the University of Oslo, Postboks 1161 Blindern, 0318 Oslo, Norway

[6]Department of Orthopedics and Orthopedic Surgery, Medical University of Graz,

Auenbruggerplatz 5, 8036 Graz, Austria

Annelie.Weinberg@t-online.de

[7]Division of Pediatric Cardiology, Johns Hopkins University School of Medicine, Baltimore,

MD, USA

wravekes@jhmi.edu

Corresponding author

Alexander Avian

Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz,

Auenbruggerplatz 2,

8036 Graz,

Austria

alexander.avian@medunigraz.at

www.medunigraz.at

Tel.: +43 – 316 -385 – 17873

Fax.: +43 – 316 -385 - 13590

What is already known about the topic

- Faces Pain Scale revised (FPS-r) has been developed as a linear interval scale.

- The scale interval scale properties of the FPS-r are questioned

What this paper adds

- Responses to the FPS-r may not be interval scaled

- When reporting responses to the FPS-r nonparametric (e.g. median, interquartile range) parameters or the number of patients above/below a certain pain level should be reported.

- Parametric parameters (e.g. mean, standard deviation) for reporting FPS-r responses may be inappropriate.

Abstract

Background

The Faces Pain Scale-revised (FPS-r) has been developed as an interval scale. For other pain measurement instruments, several studies found evidence for and against an interval level of measurement.

Objectives

The primary aim of the current study was to evaluate the scale properties of the FPS-r using an item response theory approach.

Design

Secondary analysis of published data

Setting

Three studies; Study 1 and study 2: One university hospital; Study 3: international pain registry

Participants

Study 1: n= 246, female: 41%, age: 11-18 years, 3 pain items; Study 2: n=240, female: 43%, age: 11-18 years, 9 pain items; Study 3: n=2266, female: 41%, age: 4-18 years, 3 pain items

Methods

The rating scale model (interval scale), the graded response model (no interval scale, ordered response categories) and the partial credit model (no interval scale) were used to scale the data.

Results

In all three studies, the rating scale model was outperformed by the graded response model or the partial credit model in terms of model fit. Overlapping response categories were found in items associated with less pain. Response category widths were wider for categories associated with low pain intensity and smaller for categories associated with high pain intensities. Smallest response categories were 1% to 67% smaller compared to the widest response category of the same item.

Conclusion

According to these findings, the interval scale properties of the FPS-r may be questioned. Item response theory methods may help to solve the problem of missing linearity in pain intensity ratings using FPS-r.

Contribution of the Paper statements

What is already known about the topic

- Faces Pain Scale-revised (FPS-r) has been developed as a linear interval scale.
- The interval scale properties of the FPS-r are questioned

What this paper adds

- Responses to the FPS-r cannot be assumed interval scaled
- When reporting responses to the FPS-r, nonparametric (e.g. median, interquartile range) parameters or the number of patients above/below a certain pain level should be used.
- Parametric parameters (e.g. mean, standard deviation) for reporting FPS-r responses should not be used.

INTRODUCTION

An adequate assessment tool is crucial for effective pain management in hospitals. Pain should be evaluated regularly for monitoring reasons, to follow the course of pain intensity in patients and to evaluate the effectiveness of pain therapy. Furthermore, pain is assessed for research purposes in experimental and clinical settings. Several pain assessment instruments have been established that differ in the number of items used and the target population. For children older than 4 years, the Faces Pain Scale-revised (FPS-r) is often recommended (von Baeyer, 2009).

The FPS-r has been developed as a linear interval scale. Two faces have been labeled “no pain” and “very much pain” and further four faces have been chosen in between these two faces to represent equal intervals between each of these six faces. Within the development of the FPS-r 101 different faces were presented on the computer to participants. Participants had to choose four faces that corresponded to predefined pain intensity levels on a scale with fixed endpoints (no pain to very much pain) (Hicks et al., 2001). Consequently, in many publications FPS-r values are treated as being located on an interval scale and therefore statistics relying on that scale level such as mean scores are calculated and parametric analyses are applied (e.g. Birnie et al., 2016; Brown et al., 2016; Ferreira-Valente et al., 2011; Park et al., 2015; Sánchez-Rodríguez et al., 2012). Some publications are using nonparametric analyses due to distributional concerns (e.g. Crevatin et al., 2016; , McLaughlin et al., 2016). Only in a few publications, the interval scale properties of FPS-r are questioned and therefore nonparametric analyses applied (de Azevedo et al., 2014, Hirunwiwatkul et al. 2009). Some publications even describe the FPS-r as a measurement on an ordinal scale but report parametric statistics (Ho et al., 2015). In several studies analyses of other pain measurement tools found evidence for and against an interval scale of measurements (e.g. Oliveira et al., 2014;, Shields] et al., 2003). According to von Baeyer (2009), the interval scale property of a pain measurement tool has to

be questioned even if it was explicitly designed to measure on a linear interval scale like the FPS-r. This is especially true for pain assessment in younger children (von Baeyer, 2009).

The primary aim of the current study was to evaluate the scale properties of the FPS-r and therefore assess whether the assumption that they have the properties of an interval scale holds. In three different samples responses were analyzed (e.g. response category widths, overlapping of response categories) using three different item response theory models (c.f. Box 1) for polytomous responses (Ostini and Nering, 2006). Different pain items and different age groups of patients were analyzed to examine whether the scale properties of the FPS-r are dependent upon the sets of items used and the patient's age.

Box 1 Introduction to Item Response Theory.

The basic idea of item response theory is to model item responses caused by a continuous latent variable (θ; e.g. pain intensity) (Sijtsma, 2004). To model the relationship between the probability of answering an item correctly with a person's value on the latent variable, parameters of a mathematical function are estimated. For items with more than two response categories (polytomous items), this relationship can be visualized with category response curves and category boundary curves (Pesudovs and Noble, 2005).

Important item response theory terms:

- **Category response curves** visualize the probability of responding in a certain category according to the latent trait level (θ; pain intensity) (Hays et al., 2000) (Fig. 1 left side)
- **Category boundary curves** visualize the probability to respond with a certain or a higher response category. The slope of the category boundary curves represents the ability of a response category to discriminate between patients with different pain intensities. Items with category boundary curves with very steep slopes have a high discriminating power, items with a flat slope a low discriminating power. In items with category boundary curves with very steep slopes, patients slightly below the response level threshold have a very low probability to respond with the next response category. Responding to an item with a flat slope, a patient slightly below the response level threshold would have a higher probability to answer with the next face. (Fig. 1 right side).
- **Response level thresholds** represents the value of θ where the category boundary curves reaches a probability of $p = 0.5$. For example, using the FPS-r the threshold value of the second face represents the θ value (pain intensity), where the patient has an equal chance of responding with the first or the second face (Waterman et al., 2010). A patient experiencing a higher pain intensity (θ) has a higher chance to respond with the second face. A patient experiencing a lower pain intensity has a higher chance to respond with the first face.
- **Response category widths** are the distance between two ascendant response level thresholds. A requirement for an interval scale is that the response categories have the same width. If category widths within an item are different, the amount of increase in pain intensity to respond to the next face is different for each face. While a small increase in pain intensity is sufficient for one face to respond to the next, in another face the increase in pain intensity has to be larger.

METHODS

Sample and Design

This is a secondary analysis based on three previously reported studies by our group. The rationale and design for these three studies have been reported in detail elsewhere (Avian et al., 2016, 2017). Briefly, Study 1 (Avian et al., 2016) and study 2 (Avian et al., 2017) were prospective studies (Study 1:between July 2010 and March 2012; study 2: between October 2013 and May 2014) that included patients between eleven and 18 years who underwent surgery at the Department of Pediatric and Adolescent Surgery, Medical University of Graz (Austria). Patients had to be able to speak German. Intensive care patients and patients with cognitive impairment were not included. Patients were asked by an independent researcher not involved in patient care to rate their pain-intensity.

In study 1, patients rated their pain at rest, during movement and their worst pain. This study aimed to evaluate possible order effects in children and adolescents and the possible influence of sex on order effects. Therefore, three pain items (pain at rest, during movement and their worst pain) were presented in six different orders.

Study 2 varied from the first study in that patients rated their worst pain after surgery and the pain while carrying out eight different activities. Six of these eight activities were included in this manuscript: (1) eating, (2) drinking, (3) turning over in the bed, (4) getting up from bed, (5) coughing, and (6) lying in bed. Two activities were excluded to get a unidimensional model. Study 2 aimed to analyze inconsistencies and the test-retest reliability in worst pain ratings in children and adolescents. Inconsistencies were defined as lower worst pain ratings compared to pain intensity ratings for activity pain items. In study 2, pain assessments were performed twice (t1, t2), separated by one to two hours [median time between assessments: 75 min, interquartile range (IQR): 70 – 85; Range: 60 – 120min]. In our current analysis we only included the first of the two pain assessment ratings collected in study 2.

In study 3 (Avian et al., 2017) data from an international pediatric acute pain registry (Quality Improvement in Postoperative Pain Treatment in children; QUIPSi) were included. Within the QUIPSI registry, patient data from German, Austrian and Swiss pediatric patients are collected (http://www.quips-projekt.de/). This registry includes 1) outcome measurements (pain intensity measurements, pain-related interference e.g. pain when coughing, side effects e.g. vomiting), and (2) relevant process parameters (e.g. kind of surgery, medication). Children at the age of 4 to 18 years can be included in this registry. These children were admitted for pediatric surgery in participating hospitals. These hospitals were collecting these patient data for quality improvement reasons. Within the QUIPSI registry, it is possible to compare the hospital's outcomes with all other hospital outcomes on the hospital level or e.g. on a surgery level. Of the 5970 included patients, only those answering the questionnaire alone without any help (n = 2266, 46% female, age: 13.3±2.7 years) were analyzed.

For all pain assessments, the FPS-r was used (Hicks et al., 2001). The FPS-r shows acceptable reliability in children rating their actual pain (r = .77) (Tsze et al., 2013) and moderate to high correlations with other pain assessment tools (r = .66 - .87) (Park et al., 2015; Tsze et al., 2013).

Ethical considerations

All three studies comply with all institutional guidelines related to patient confidentiality and research ethics including institutional review board approval (Study 1 and 2: Medical University Graz Ethics Committee, IRB00002556; Study 3: University Ethics committee of Jena University Hospital, Thuringia, Germany, IRB00004153).

Data analysis

The data sets of the three studies were analyzed separately. The R-package mirt (version 1.25) (Chalmers, 2012, 2017) was used for data analysis. The software R (Version 3.4.1, 2017-06-30; R Foundation for statistical Computing) was used for all analyses. Missing data were not imputed. Response categories were collapsed if less than 10 responses within a category were observed. To analyze the scale properties of the polytomous pain ratings, three different item

response theory models were compared: the rating scale model (Andrich, 1978), the graded response model (Samejima, 1969) and the partial credit model (Masters, 1982).

While the rating scale model assumes an interval scale, the graded response model only assumes ordinal scale and the partial credit model assumes distinct responses, which do not have to be ordered. For an interval scale, the rating scale model has to have a better model fit compared to graded response model and partial credit model, and for an ordinal scale, the graded response model a better model fit compared to partial credit model.

The rating scale model (Andrich, 1978) was developed to estimate latent traits (θ) using the response patterns on items with polytomous answering modes. In this study the latent trait to be estimated is the underlying pain intensity. The rating scale model models the response categories with equidistant thresholds. Therefore the probability for a specific response depends on the two threshold parameters m (number of item thresholds) and l (number of thresholds before selected response category), the two location parameters b_i (item location parameter) and τ_g (boundary between the categories relative to each item's trait location).

The graded response model (Samejima, 1969) assumes more than two ordinal response categories. A higher value in the latent trait θ (here: pain intensity) goes along with a higher score in the response category. Dependent on the person's θ, it models the likelihood of a person using a certain response category or a higher response category. Within an item i all response categories have the same discrimination parameter a_i but separate difficulty parameters b_{ig} for each response category g. The index i indicates that the discrimination parameters can also differ between items. Each response category is defined by a_i and the difficulty parameter b_{ig}.

Within the partial credit model, threshold parameters do not necessarily need to be ordered. Therefore, response categories do not need to be ordered (Masters, 1982). This leads to a different interpretation of boundary parameters in the graded response model and the partial credit model.

Different item response theory models differ in the way the likelihood for a person with a given latent trait θ to respond with a certain response is calculated (Box 2). Therefore, the interpretations of these thresholds differ slightly. While the graded response model assumes ordered response categories and therefore opposes all responses below a certain threshold to all responses above this threshold, the partial credit model always opposes two response categories, which are supposed to be next to each other, without assuming an order.

According to the different model assumptions of the used item response models category response curves and category boundary curves differ. Consequently, response categories are not necessarily ordered in partial credit model and therefore category response curves and category boundary curves are not ordered. Furthermore, the distance between two ascendant category boundary curves are only equal in rating scale models (Fig. 1).

Box 2: The likelihood for a person with a given latent trait θ to respond with a certain response according to three different item response models (Ostini and Nering, 2006).

Rating scale model:

$$P_{i_g} = \frac{\exp^{\sum_{g=0}^{l}[\theta-(b_i-\tau_g)]}}{\sum_{h=0}^{m}\exp^{\sum_{g=0}^{h}[\theta-(b_i-\tau_g)]}}$$

Graded response model

$$P_{i_g} = \frac{\exp^{a_i\left(\theta-b_{i_g}\right)}}{1+\exp^{a_i\left(\theta-b_{i_g}\right)}}$$

Partial credit model

$$P_{i_g} = \frac{\exp^{\sum_{g=0}^{l}\left(\theta-b_{i_g}\right)}}{\sum_{h=0}^{m}\exp^{\sum_{g=0}^{h}\left(\theta-b_{i_g}\right)}}$$

θ …latent trait (e.g. overall pain level)

g … certain response (e.g. a certain face of the FPS-r, or a certain number of a numeric rating scale)

h … all possible responses. The possible responses range from 0 to *m*. *g* is one element of *h*.

i … answered item

P_{ig} ...the likelihood for a person with a given latent trait θ to respond to an item *i* with a response *g*.

m … number of item thresholds

l … number of thresholds before selected response category

b_i … item location parameter

τ_g … boundary between the categories relative to each item's trait location

a_i … discrimination parameter

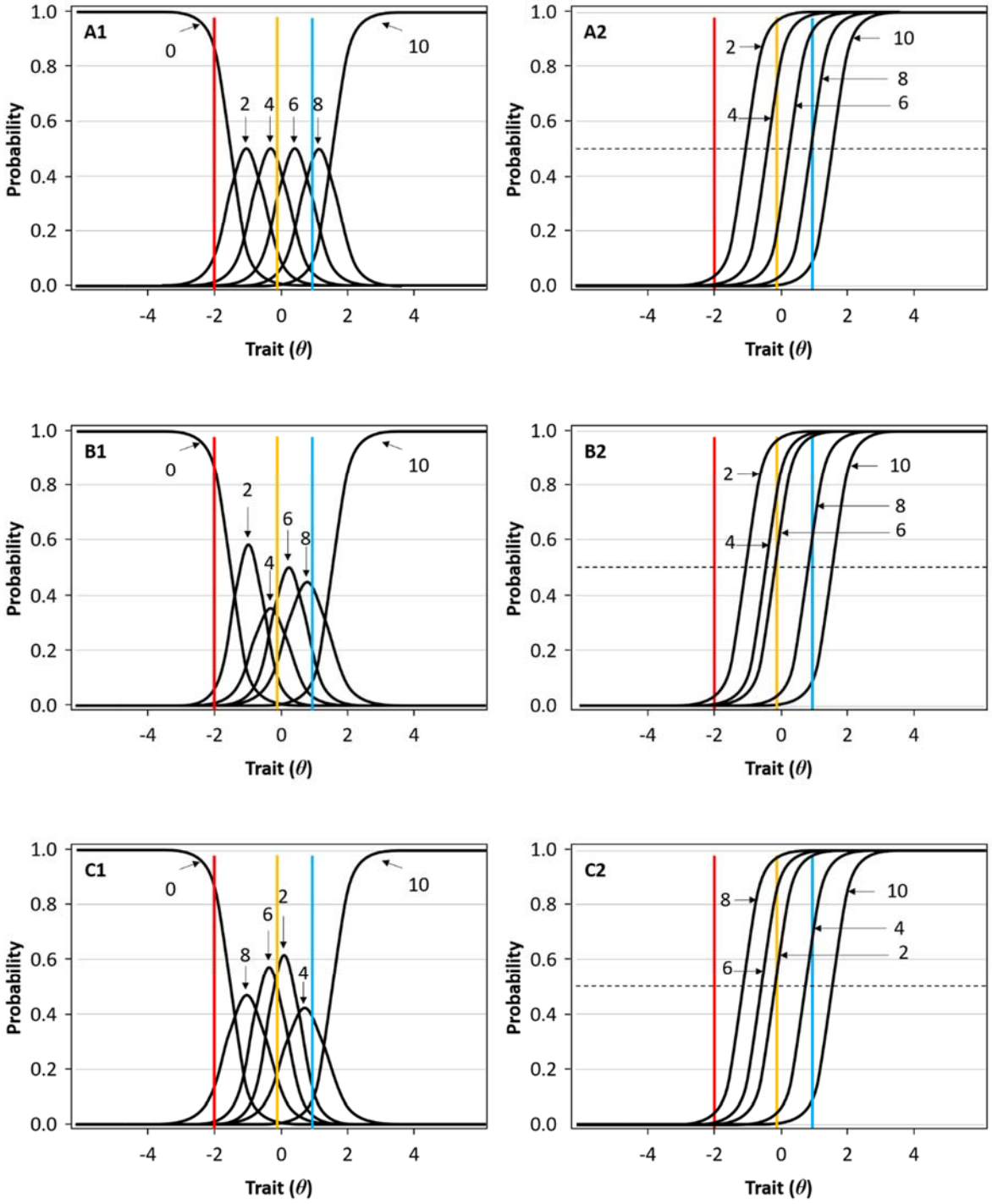


Figure 1 Examples for category response curves (left side) and category boundary curves (right side) for (A) the rating scale model, (B) the graded response model and (C) the partial credit model.

Footnote: In each figure, three possible patients (red, yellow and blue) are marked. The red patient has low pain, the yellow intermediate and the blue high pain. Response categories are marked with 0, 2, 4, 6, 8 and 10

indicating the response category's pain intensity. Using the category response curve, the likelihood for a person with a given latent trait θ to respond with a certain response (P_{ig}; cf. Box 2) can be seen. For all three models, the red patient has the highest probability to respond with "0". The yellow patient has the highest probability for the ratings scale model, the graded response model and the partial credit model to respond with 4, 6 and 2 respectively. The blue patient has the highest probability to respond with 8 for the rating scale method and the graded response model. For the partial credit model, the highest probability is to respond with 4. The characteristics of the three models can be seen in both figures. In the rating scale model the category response curves are ordered and have the same height. The category boundary curves are also ordered and the distance between each curve is the same. In the graded response model the category response curves are also ordered. The height and width of the curves are different. Therefore, the distances of the category boundary curves are different. In the partial credit model, the category response curves and the category boundary curves are disordered.

The three models were compared using likelihood-based statistics. Fit indices (AIC: Akaike information criterion, BIC: Bayesian information criterion, AICc: AIC corrected for small sample sizes, SABIC: sample size adjusted BIC) (Box 3) were used to decide which model showed the better model fit. Since these fit indices are relative fit indices, there is no absolute bound that can be used to decide whether a model fits or does not fit. Therefore, the comparison of two fit indices is used to decide which model fits better. A lower fit index indicates a better model fit. For the calculation of AIC, AICc, BIC and SABIC, the maximum likelihood ($\hat{L}$) of each model is used. The fit indices differ in the penalty weight that is added to the term -2ln ($\hat{L}$). The penalty weights depend on the number of estimated parameters (degrees of freedom: *df*). Furthermore all penalty weights except for AIC depend on the sample size (n) (Stoics and Selén, 2004). The partial credit model is a more general form of the rating scale model. Apart from the parameter τ_g both models are the same. Therefore, these models are called nested models and can be compared using a likelihood ratio test (c.f. Box 2).

Box 3 Calculation of fit indices (Stoica and Selén, 2004).

$\hat{L}$ … maximum value of the likelihood function of the model

n … sample size

df … degrees of freedom (number of parameters estimated by the model)

$$AIC = -2 * ln(\hat{L}) + 2 * df$$

$$AICc = -2 * \ln(\hat{L}) + 2 * \frac{n}{n - df - 1}$$

$$BIC = -2 * \ln(\hat{L}) + df * \ln(n)$$

$$SABIC = -2 * \ln(\hat{L}) + df * \ln\left(\frac{n + 2}{24}\right)$$

AIC … Akaike information criterion

AICc … AIC corrected for small sample sizes

BIC … Bayesian information criterion

SABIC … sample size adjusted BIC

For the final model test, category response curves and category boundary curves for all items were produced and response level threshold values for all items were calculated. In Study 3 category response curve, category boundary curves and response level thresholds were calculated for (1) all patients and (2) separately for age groups that have been predefined within the registry (4-9 years, 10-12 years, 13-14 years, 15-16 years and 17-18 years).

RESULTS

Sample characteristics

All patients of the original analysis of study 1 and study 2 were also included in this secondary analysis. Out of the 5970 patients included in the original analysis of study 3 only

those 2266 answering the questionnaire alone without any help were included. The characteristics of these three samples are given in Table 1.

Table 1 Characteristics of analyzed patients.

	Study 1 (Avian et al. 2016)	Study 2 (Avian et al. 2017)	Study 3 (Avian et al. 2017)
n	246	240	2266
age (years), mean±SD	14.4±2.0	14.7±1.9	13.3±2.7
female/male, n	101/145	103/137	1041/1200 missing: 25
duration of surgery (min), median (IQR)	37 (21-68)	53 (33 – 95)	46 (25 – 84)
worst pain, median (IQR)	4 (2-6)	4 (2-6)	4 (2-6)

Footnote:

SD … standard deviation

IQR … interquartile range

Fit indices

According to the fit indices, the graded response model showed the best model fit for all three studies. The information criteria AIC, AICc and SABIC for the graded response model were lowest in all studies, and BIC in study 3. In study 1 and 2 BIC was higher for the graded response model (study 1 BIC: 1842.6; study 2 BIC: 3160.1) compared to rating scale model (study 1 BIC: 1817.2; study 2 BIC: 3155.1) and partial credit model (study 1 BIC: 1848.7;study 2 BIC: 3153.4). The partial credit model outperformed the rating scale model only in study 2 (χ^2=67.4, df=12, p <.001). Based on the results regarding model fit, for the calculation of category boundary curves, category response curves, category thresholds, and category widths, the graded response model was used for all studies (Table 2).

Table 2 Fit indices and model comparison for all analyzed studies

		AIC	AICc	BIC	SABIC	$\ln(\hat{L})$	Number of estimated parameters	RSM vs. PCM sign. (df; χ^2)
Study 1	RSM	1789.1	1789.7	1817.2	1791.8	-886.6	8	.131 (8; 12.5)
	GRM	1779.5	1782.5	1842.6	1785.5	-871.7	18	
	PCM	1792.6	1795.0	1848.7	1798.0	-880.3	16	
Study 2	RSM	3120.3	3121.2	3155.1	3123.4	-1550.1	10	<.001 (12; 67.4)
	GRM	3062.7	3070.4	3160.1	3071.4	-1503.3	28	
	PCM	3076.9	3081.5	3153.4	3083.7	-1516.4	22	
Study 3	RSM	17107.5	17107.6	17153.3	17127.9	-8545.7	8	.052 (8; 15.4)
	GRM	16936.7	16937.0	17039.7	16982.5	-8450.3	18	
	PCM	17108.1	17108.4	17199.7	17148.9	-8538.1	16	

RSM … rating scale model

GRM … graded response model

PCM … partial credit model

AIC … Akaike information criterion

AICc … AIC corrected for small sample sizes

BIC … Bayesian information criterion

$\hat{L}$ … maximum value of the likelihood function of the model

df … degrees of freedom

χ^2 … Chi square value

Category boundary curves, category thresholds and category widths

In figure 2 category boundary curves are shown for the whole group of study 1 (category boundary curves for study 2 and study 3 are given in the supplemental figures 1A – 1B). Response level threshold values, item slopes, category widths and the reduction in category widths are given for each item of study 1, study 2, and study 3 (whole group and age groups) in Table 3.

The slopes associated with each response level range from 1.20 to 4.58. Lower slopes ($a \leq 2.0$) were found in four activity pain items (getting up from bed, drinking, eating, and coughing) in study 2 (Table 3). These items have a lower ability to discriminate between patients experiencing low and patients experiencing higher pain intensities. Intermediate slopes ($2.0 < a \leq 3.0$) were found in younger children (4 – 9 years), for pain at rest item in study 3, movement pain in study 1 and two items in study 2 (worst pain, lying in bed). Items with the best ability to discriminate patients experiencing low and patients experiencing higher pain intensities (slopes: $a > 3.0$) were observed for worst pain and pain at rest in study 1, one activity pain rating of study 2 (turning over in the bed), movement pain and worst pain in study 3 (except 4-9 year old children).

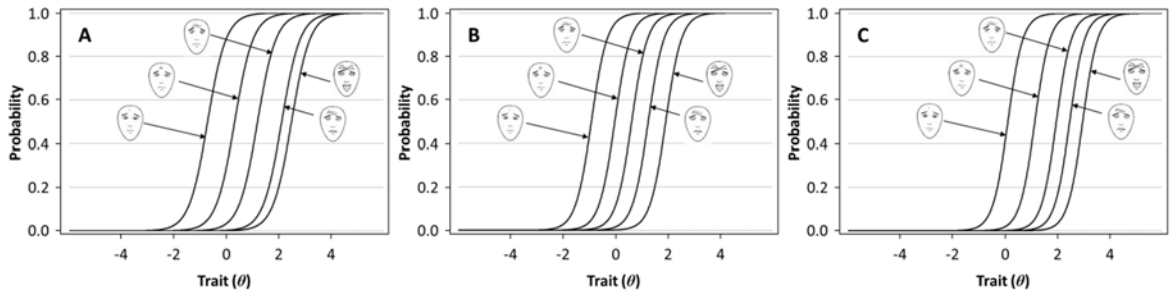


Figure 2 Category boundary curves for movement pain (A), worst pain (B) and pain at rest (C) in study 1 (n = 246) using the graded response model.

As can be seen in Figure 2 and in Table 3, category widths of response categories are wider for low pain intensity and smaller for high pain intensities. In study 1 and study 3 the widest category was always found between the 1st and 2nd threshold (Figure 2, Supplemental Figure 1c, Table 3). In study 2 the same result was found in five out of six items. Only in the activity pain item "eating", the smallest category was found between the 1st and 2nd threshold (Supplemental Figure 1b. Table 3). In all items except one item response categories were ordered correctly. For the pain item "cough" in study 2, response categories were out of order (Table 3). Therefore, the category width was not analyzed. In eleven out of 25 analyzed items, the smallest category width was found in the second highest response category. Since the width of a category is defined as the distance between two thresholds, no width is given for the lowest and highest response category. Therefore, the second highest response category is the response category associated with the highest pain intensity for which a category width has been calculated. Since categories have been collapsed, these second highest response categories were four times the 5th response category, two times the 4th response category (5th and 6th response categories were collapsed) and five times the 3rd response category (4th, 5th and 6th response categories were collapsed). In further twelve items the smallest category width was in the third highest response category. The third highest response category was found eleven times in the 4th response category and one time in the 3rd response category. Comparing smallest and widest response categories within each item, the biggest differences were found in children older than 13 years answering pain at rest items (study 3), in study 1 for movement pain and study 2 for worst pain (difference in width: 61.2% - 66.9%). The smallest difference was found in the "drinking" item of study 2 (difference in width: 1.5%) (Table 3).

Category response curves

Looking at the response categories with the highest probability to be chosen for a specific pain level, some response categories never have the highest probability. These response categories were overlapped by other response categories. Overlapping response categories were found in

one item in study 1 (pain at rest) (Figure 3), one item in study 2 (cough) (figure 4) and in children ≥ 13 years for pain at rest (Supplemental figure 2). Analyzing all age groups of study 3 together, no overlapping response categories were found (figure 5). In three out of these four items (study 1: pain at rest; study 3: pain at rest in 13 – 14 year old patients; study 3: pain at rest in 15 – 18 year old patients) the overlapped response category was the second highest response category. In the fourth item (study 2: cough) the lowest and the highest response category overlapped all other response categories. Some response categories have only small ranges (<0.5) where they have the highest probability to be chosen. These kind of response categories were found in study 1 in one item (pain at rest, second and third highest response category) (Figure 3) and in study 3 in young children (age <10 years: movement pain, worst pain; age 10-11: movement pain, pain at rest) (Supplemental figure 2). Similar to the overlapping response categories, these small categories were found in response categories associated with higher pain intensity (second and third highest response categories). All category response curves are given in figure 3 – 5. For study 3 the response categories for the different age groups are given in the supplemental figure 2.

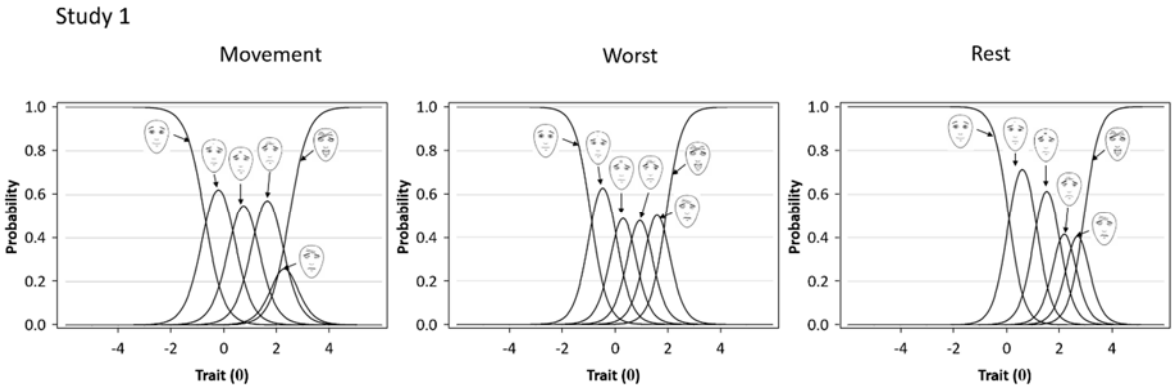


Figure 3 Category response curves for movement pain, worst pain and pain at rest in study 1 (n = 246) using the graded response model.

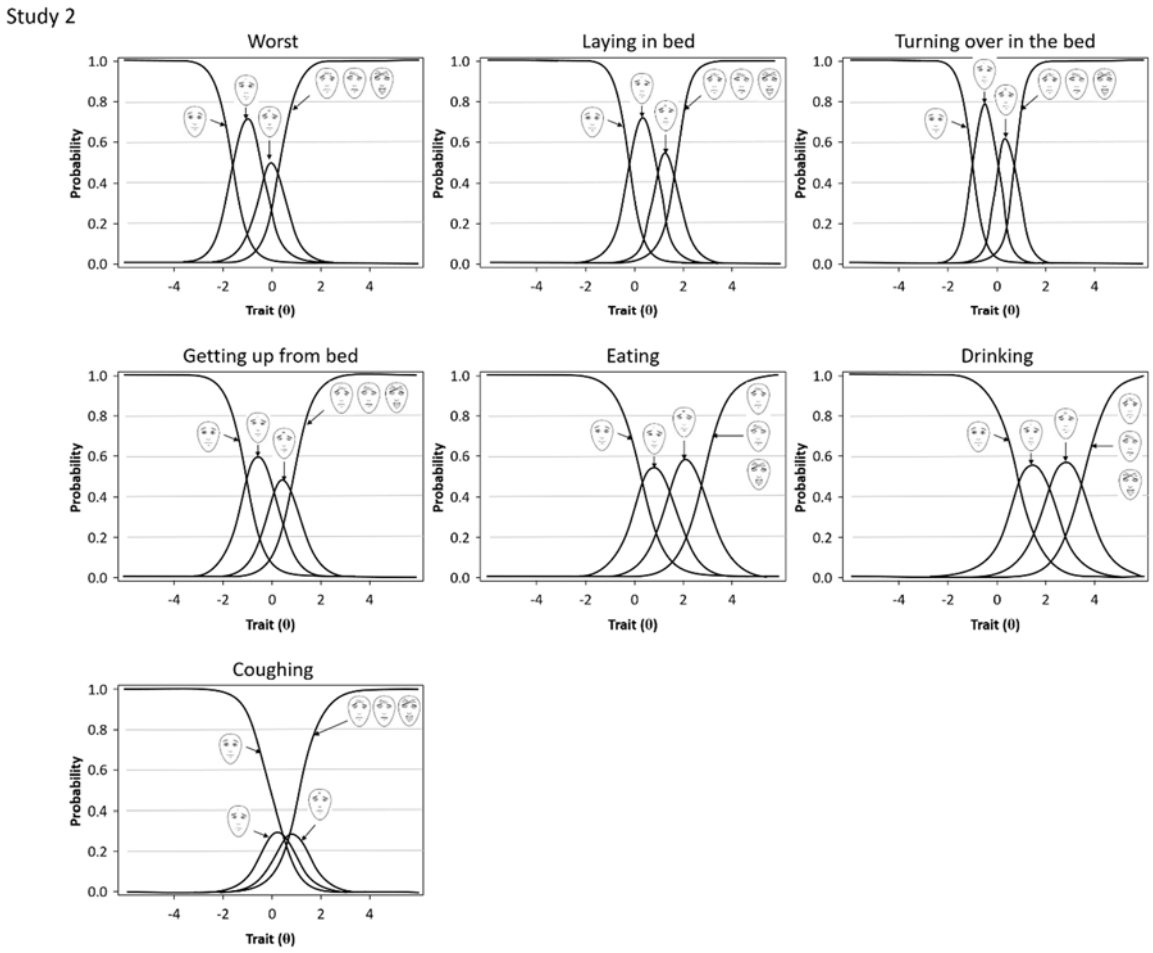


Figure 4 Category response curves for worst pain and activity pain items in study 2 (n = 240) using the graded response model.

Footnote: Categories 4 to 6 were collapsed due to the small number of patients in single response categories.

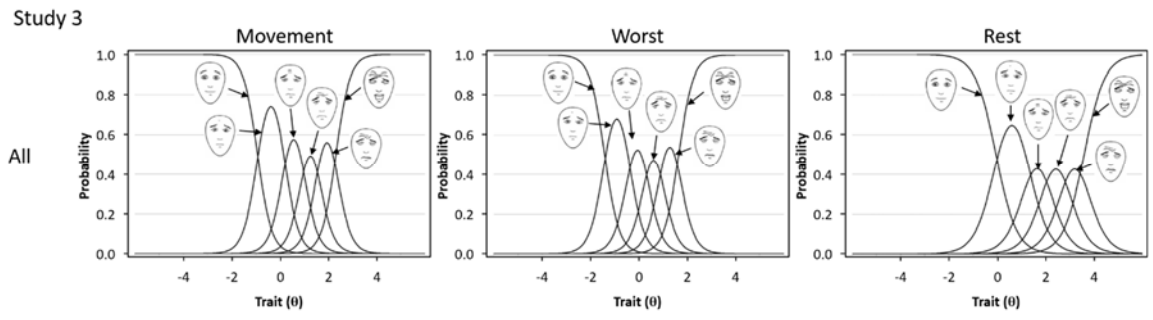


Figure 5 Category response curves for movement pain, worst pain and pain at rest in study 3 (n = 2266) using the graded response model.

Table 3 Slopes, category thresholds and category width for the items of the analyzed studies

	n	Slope	Category threshold					Category width						Difference max-min	
		a	b1	b2	b3	b4	b5	c2	c3	c4	c5	Min	Max	absolute reduction	relative reduction
Study 1															
Movement	246	2.81	-0.71	0.32	1.20	2.12	2.50	1.03	0.88	0.92	0.38	0.38	1.03	0.66	63.6%
Worst	246	3.30	-0.91	0.01	0.64	1.27	1.94	0.92	0.63	0.63	0.67	0.63	0.92	0.29	31.3%
At rest	246	3.46	0.11	1.14	1.97	2.48	2.98	1.03	0.82	0.51	0.50	0.50	1.03	0.54	51.8%
Study 2[1]															
worst	240	2.23	-1.68	-0.28	0.23			1.40	0.50			0.50	1.40	0.90	64.1%
getting up from bed	240	1.84	-1.03	0.19	0.72			1.22	0.53			0.53	1.22	0.68	56.1%
turning over in the bed	240	3.71	-0.99	0.06	0.70			1.05	0.65			0.65	1.05	0.40	38.0%
coughing	240	1.20	0.58	0.57	0.54			-0.02	-0.02			[2]			
lying in bed	240	2.82	-0.20	0.93	1.55			1.13	0.62			0.62	1.13	0.51	45.0%
eating	240	1.55	0.42	1.39	2.59			0.96	1.20			0.96	1.20	0.24	19.8%
drinking	240	1.61	1.00	2.02	3.02			1.01	1.00			1.00	1.01	0.01	1.5%
Study 3															
Movement															
All	2266[3]	3.42	-0.93	0.19	0.96	1.59	2.32	1.12	0.77	0.63	0.74	0.63	1.12	0.49	43.6%
4-9 years	269	2.40	-0.61	0.57	1.40	2.01		1.18	0.83	0.61		0.61	1.18	0.57	48.4%
10 - 12 years	693	3.14	-0.88	0.25	0.98	1.58	2.28	1.13	0.72	0.61	0.69	0.61	1.13	0.52	46.0%
13 - 14 years	614	4.58	-0.96	0.12	0.86	1.43	2.22	1.08	0.75	0.57	0.78	0.57	1.08	0.51	47.2%

15 - 18 years	649	3.41	-1.09	0.08	0.90	1.59	2.36	1.17	0.82	0.70	0.77	0.70	1.17	0.47	40.5%
Worst															
All	2266[1]	3.28	-1.42	-0.41	0.30	0.92	1.65	1.01	0.71	0.62	0.73	0.62	1.01	0.39	38.3%
4-9 years	269	2.66	-1.02	-0.03	0.63	1.11		0.99	0.66	0.48		0.48	0.99	0.51	51.9%
10 - 12 years	693	3.33	-1.38	-0.38	0.35	0.93	1.62	0.99	0.74	0.57	0.70	0.57	0.99	0.42	42.4%
13 - 14 years	614	3.66	-1.40	-0.46	0.22	0.87	1.68	0.93	0.68	0.65	0.82	0.65	0.93	0.28	30.3%
15 - 18 years	649	3.41	-1.69	-0.51	0.20	0.89	1.62	1.18	0.71	0.69	0.73	0.69	1.18	0.49	41.7%
At rest															
All	2266[2]	2.34	-0.05	1.27	2.05	2.83	3.61	1.32	0.78	0.78	0.78	0.78	1.32	0.54	40.9%
4-9 years	269	2.14	0.35	1.55	2.35	3.26		1.20	0.80	0.92		0.80	1.20	0.40	33.3%
10 - 12 years	693	2.32	0.01	1.27	1.97	2.59	3.59	1.26	0.70	0.62	1.01	0.62	1.26	0.65	51.3%
13 - 14 years	614	2.54	-0.12	1.14	1.99	2.89	3.38	1.26	0.85	0.90	0.49	0.49	1.26	0.77	61.2%
15 - 18 years	649	2.15	-0.19	1.33	2.18	3.18	3.68	1.52	0.85	1.00	0.50	0.50	1.52	1.01	66.9%

[1] Categories 4 to 6 were collapsed due to the small number of patients in single response categories.
[2] categories out of order; therefore no min or max was computed.
[3] in 41 patients age is missing
a … Slope of each item
b1 – b5 … Category threshold. b1 refers to the threshold between the first and the second face, b2 between the second and the third face a.s.o
c2 – c5 … Category width. c2 refers to the category width of the second face, c3 to the category width of the third face a.s.o. Since the first face do not have a lower and the last face do not have an upper threshold, no category width were calculated for these faces.

DISCUSSION

The main finding was the lack of interval scale property of the FPS-r across all analyzed studies. Since three relatively large data sets from different studies were analyzed, there is strong evidence that the interval scale assumption is generally inappropriate for the FPS-r. Overlapping response categories or response categories covering smaller ranges were primarily found in higher response categories of items associated with less pain (pain at rest, cough). An interpretation of these response categories is therefore limited. In older pediatric patients, overlapping response categories/categories covering small ranges were found in one out of three, and in younger patients in two out of three items.

Regardless of age or kind of pain item, the widest category width was mainly found in the second response category (pain intensity = 2). In our institution this category is interpreted as having low pain and no pain medication is necessary (Messerer et al. 2010). The smallest categories were those associated with the 2nd or 3rd highest pain intensity. Pesudovs and Noble (2005) also found different response category widths using a faces pain scale with small differences between the 2nd and 3rd category and between 4th and 5th category.

The overlapping response categories/categories covering small ranges in younger children might be caused by difficulties in distinguishing six different response categories. Using item response theory methods, Pesudovs and Noble (2005) analyzed a 7-category faces pain scale in adults. They found that response category 5 was underused resulting in overlapping response categories. After collapsing two response categories, a linear measurement on a continuous latent variable could be achieved. Decruynaere et al. (2009) could show that young children (4-7 years) are not able to distinguish more than two (4-5 years) or three (6-7 years) faces of a face scale. In line with this finding, Hamilton (1968) reported already in the year 1968 in his review strong evidence for a preference of extreme responses in children and adolescents compared to adults. Standford et al. (2006) could also observe an age dependency in the ability of using the FPS-r. The accuracy of pain intensity ratings continuously improved from 3 to 6-year old

children. Nevertheless, half of the 6-year old children had difficulties using the FPS-r. Apart from age, no other factor could be identified that influences the accuracy of pain intensity ratings. Champion et al. (1998) previously showed that children's cognitive maturity is a crucial factor for applying self-report scales.

According to our results and those from the literature, children and adolescents do not use the FPS-r in an interval manner. In other pain assessment tools, the comparability of intervals is also questioned. After studying the pain assessments of children using the Visual analogue Scale (VAS), Berntson and Svensson (2001) conclude that the VAS does not have interval scale properties. Similar results were found by Shields et al. (2003). They relate the ability to use the VAS correctly to the ability of abstract thinking. Contrary to these findings, Myles et al. (1999) could show ratio scale properties of the VAS in adult patients at least for mild-to moderate pain. Comparing numeric rating scale (NRS) values to VAS values in adults, Hartrick et al. (2003) conclude that NRS shows ratio scale properties only in certain situations (e.g. laboring patients). Oliveira et al. (2014) conclude for the Wong-Baker FACES Pain Rating Scale that in children up to an age of 8 years it does not have interval scale properties. For older children interval scale properties may be met if these children have a history of chronic pain. Von Baeyer (2009) already pointed out that the question whether a pain measurement leads to interval or to ordinal quality measurements is a question depending not only on the measurement tool but also on patients' characteristics (e.g. age).

If pain measurements in young children do not have at least interval scale properties, important restrictions for research and clinical routine have to be made. Calculating a difference of pain intensity ratings cannot be interpreted in a meaningful way anymore. In the literature, the minimum clinically important difference when using the VAS in children was found to be 10 mm (Powell et al., 2001). In this study they used the criteria of "a bit better". Using another criterion ("a lot better"), Kelly (2001) came up with a minimum clinically important difference of 20 mm. If response category widths are different depending on the pain intensity, these fixed

intervals in pain ratings are associated with different changes in real pain intensity and are thus lacking a solid basis for a valid interpretation. Other approaches in meaningful outcomes are proposed by Moore et al. (2010). For chronic pain they suggest an improvement of 30% (moderate benefit) or 50% (substantial benefit), proportion of patients below 30/100 mm or patients' global impression (very much improved). If smaller response categories are always associated with higher pain intensities, then using reduction in percent would help to overcome the problem of smaller response category. Defining the targeted outcome in proportion of patients below a certain cut off point has the advantage of requiring only ordinal scaled pain scores. While this approach is fine for group level assessment, it faces problems in applying it to daily routine work.

Like every other measurement, pain measurement is faced with measurement errors. Therefore, reported pain scores can be viewed as a good estimation of the real pain intensity, but real pain intensity may be a little bit higher or lower. It could be shown that the limits of agreement derived from a Bland-Altman Plot (Altman and Bland, 1983) of two consecutive pain assessments using FPS-r are in the area of ±1.5/10 to ±2.6/10, depending on the kind of pain that has been assessed (e.g. worst pain, pain caused by cough a.s.o) (Avian et al. 2017). Other pain assessment tools have shown slightly lower limits of agreement (±1.1/10 to ±1.9/10) (Bailey et al., 2010, 2012).

If more than one pain item is used to calculate a composite pain score, item response theory methods could be used for pain intensity estimation. Using item response theory methods, the pain intensity is estimated using not only patients' response but also item specific parameters (e.g. what pain intensity is associated with a certain pain item). Therefore, these item specific parameters have to be calculated using a representative sample of patients. While this approach will result in linear scores for pain intensity (Pesudovs and Noble, 2005), it has to be considered that calculating individual pain scores will be more complicated and interpretation of these

scores will be challenging. In item response theory, individual pain scores cannot be calculated by summing up the responses but are derived using specific formulas. Therefore, electronic devices have to be used. Legal requirements (e.g. medical devices acts) have to be considered if formulas are implemented in electronic devices for patients' care purposes. Furthermore, the resulting pain intensity score will not be in the range of 0–10. Even if these scores are transformed to a familiar 0-10 scale, they do not necessarily have comparable properties like established pain scales (e.g. comparable cut off scores).

This study had limitations that future research should address. We included only a small number of young children and therefore grouped these children in one age group (age <10 years). This small number of young children is due to the fact that most of the children in this age group answered the questionnaire with the aid of their parents (71%) or someone else did the pain ratings (20%). Since we aimed to get an impression of children's ability to answer pain scales, we did not include these children. To get a better understanding of these children's abilities, further research is necessary.

In two out of the three analyzed studies, only three pain items were assessed. In item response theory the target outcome of each patient is estimated according to response patterns to the answered items. Three items are a very small number for this estimation process. Therefore, the results regarding these two studies have to be interpreted with caution. Nevertheless, the main results are comparable in all three studies. The age group of <10 years could only be analyzed in study 3.

One further limitation is that this is a secondary analysis of data collected for another purpose. Therefore, these results should be verified in prospective studies. However, the analysis of three different studies, which differ in many ways, may strengthen the conclusion since similar results were found in different settings.

In this study three different item response theory models were compared. Aside from these models several other models are available, e.g. for a less restrictive situation (not equally spaced

response categories and not ordered response categories) the generalized partial credit model (including a discriminatory parameter) (Muraki, 1992) or the nominal response model (Bock, 1072). Comparing other models may lead to other conclusions.

CONCLUSION

The main findings of our study are wider response category widths for low pain intensity categories compared to higher pain intensity categories and overlapping response categories especially in younger children and pain items with lower pain intensities. According to these findings the interval scale properties of the FPS-r may be questioned. Item response theory methods may help to solve the problem of missing linearity in pain intensity ratings using FPS-r.

ACKNOWLEDGMENTS

The authors wish to thank C. von Baeyer for his valuable comments during the preparation of this manuscript. The authors have no conflicts of interest to declare. This study was supported by funds of the Oesterreichische Nationalbank (Oesterreichische Nationalbank, Anniversary Fund, project number: 14335) and the DFG (German Research Foundation, project number FR 2552/5-1). The first author was supported by the DFG and Oesterreichische Nationalbank. The funding source had no role in the study design; data collection, analysis, or interpretation of the data or writing of the report. The authors thank C. Weinmann for her excellent English proofreading.

References

1. Altman, D.G., Bland, J.M., 1983. Measurement in Medicine: the Analysis of Method Comparison Studies. The Statistician 32 (1983) 307-317.
2. Andrich, D., 1978. A rating scale formulation for ordered response categories. Psychometrika. 43, 561-573.
3. Avian, A., Messerer, B., Weinberg, A., Meissner, W., Schneider, C., Berghold, A., 2016. The impact of item order and sex on pain expression in children and adolescents. Health Psychol. 35, 483-491.
4. Avian, A., Messerer, B., Meissner, W., Sandner-Kiesling, A., Kammel, J., Labugger, M., Weinberg, A., Berghold, A., 2017. Using a worst pain intensity measure in children and adolescents. Journal of Advanced Nursing. 73, 1873-1883.
5. Bailey, B., Daoust, R., Doyon-Trottier, E., Dauphin-Pierre, S., Gravel, J., 2010. Validation and properties of the verbal numeric scale in children with acute pain. Pain. 149, 216-221.
6. Bailey, B., Gravel, J., Daoust, R., 2012. Reliability of the visual analog scale in children with acute pain in the emergency department. Pain. 153, 839-842.
7. Bernston, L., Svensson, E., 2001. Pain assessment in children with juvenile chronis arthritis: a matter of scaling and rater. Acta Paediatr. 90, 1131–1136.
8. Birnie, K.A., Chambers, C.T., Chorney, J., Fernandez, C.V., McGrath, P.J., 2016. Dyadic analysis of child and parent trait and state pain catastrophizing in the process of children's pain communication. Pain. 157, 938–948.
9. Bock, R. D., 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 37, 29-51.
10. Brown, R., Fortier, M.A., Zolghadr, S., Gulur, P., Jenkins, B.N., Kain, Z.N., 2016. Postoperative Pain Management in Children of Hispanic Origin: A Descriptive Cohort Study. Anesth Analg. 122, 497-502.

11. Chalmers, R.P., 2012. mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. 48, 1-29.
12. Chalmers R.P., 2017. The mirt package: multidimensional item response theory. Library of the R package; www.cran.r-project.org/web/packages/mirt/mirt.pdf. (accessed 1 August 2017)
13. Champion ,G.D., Goodenough, B., von Baeyer, C.L., Thomas, W., 1998. Measurement of pain by self-report, in: Finley G.A., McGrath P.J. (Eds.), Measurement of pain in infants and children. IASP Press, Seattle WA, pp. 123–60.
14. Crevatin, F., Cozzi, G., Braido, E., Bertossa, G., Rizzitelli, P., Lionetti, D., Matassi, D., Calusa, D., Ronfani, L., Barbi, E., 2016. Hand-held computers can help to distract children undergoing painful venipuncture procedures. Acta Paediatr. 105, 930-934.
15. de Azevedo, C.B., Carenzi, L.R., de Queiroz, D.L.C., Anselmo-Lima, W.T., Valera, F.C.P., Tamashiro, E., 2014. Clinical utility of PPPM and FPS-R to quantify post-tonsillectomy pain in children. Int J Pediatr Otorhinolaryngol. 78, 296–299.
16. Decruynaere, C., Thonnard, J.L., Plaghki, L., 2009. How many response levels do children distinguish on faces scales for pain assessment? Eur J Pain. 13, 641-648.
17. Ferreira-Valente, M.A., Pais-Ribeiro, J.L., Jensen, M.P., 2011. Validity of four pain intensity rating scales. Pain. 152, 2399–2404.
18. Hamilton, D. L., 1968. Personality Attributes Associated with Extreme Response Style. Psychol Bull. 69, 192-203.
19. Hartrick, C.T., Kovan, J.P., Shapiro, S., 2003. The Numeric Rating Scale for Clinical Pain Measurement: A Ratio Measure? Pain Pract. 3, 310–316.
20. Hays, R.D., Morales, L.S., Reise, S.P., 2000. Item Response Theory and Health Outcome Measurement in the 21sr Century. Med Care. 38, 1128-1142.

21. Hicks, C.L., von Baeyer, C.L., Spafford, P.A., van Korlaar, I., Goodenough, B., 2001. The Faces Pain Scale - Revised: toward a common metric in pediatric pain measurement. Pain. 93, 173-183.
22. Hirunwiwatkul, P., Chaikitthai, N., Wadwongtham, W., Tansatit, T., 2009. Vessel sealing system tonsillectomy vs cold knife tonsillectomy: a randomized, paired control study of efficacy and adverse effects. Asian Biomedicine. 3, 487-495.
23. Ho. E.S., Curtis. C.G., Clarke. H.M., 2015. Pain in Children Following Microsurgical Reconstruction for Obstetrical Brachial Plexus Palsy. J Hand Surg Am. 40, 1177-1183.
24. Kelly, A.M., 2001. Setting the benchmark for research in the management of acute pain in emergency departments. Emerg Med (Fremantle). 13, 57–60.
25. Masters, G.N., 1982. Rasch model for partial credit scoring. Psychometrika. 47, 851-859.
26. McLaughlin, J.M., Lambing, A., Witkop, M.L., Anderson, T.L., Munn, J., Tortella, B., 2016. Racial Differences in Chronic Pain and Quality of Life among Adolescents and Young Adults with Moderate or Severe Hemophilia. J Racial Ethn Health Disparities. 3, 11–20.
27. Messerer, B., Gutmann, A., Weinberg, A., Sandner-Kiesling, A., 2010. Implementation of a standardized pain management in a pediatric surgery unit. Pediatr Surg Int. 26, 879-889.
28. Moore, R.A., Eccleston, C., Derry, S., Wiffen, P., Bell, R.F., Straube, S., McQuay, H. for the ACTINPAIN Writing Group of the IASP Special Interest Group (SIG) on Systematic Reviews in Pain Relief and the Cochrane Pain, Palliative and Supportive Care Systematic Review Group Editors, 2010. “Evidence” in chronic pain: establishing best practice in the reporting of systematic reviews. Pain.150, 386–389.

29. Muraki, E.,1992. A generalized partial credit model: application of an EM algorithm. Applied Psychological Measurement. 16, 159 - 176.

30. Myles, P.S., Troedel, S., Boquest, M., Reeves, M., 1999. The Pain Visual Analog Scale: Is It Linear or Nonlinear? Anesth Analg. 89, 1517–1520.

31. Oliveira, A.M., Batalha, L.M.C., Fernandes, A.M., Gonçalves, J.C., Viegas, R.G., 2014. A functional analysis of the Wong-Baker Faces Pain Rating Scale: linearity, discriminability and amplitude. Revista de Enfermagem Referência. 3, 121-130.

32. Ostini, R., Nering, M.L., 2006. Polytomous item response theory models. Sage Publications, Thousand Oaks CA.

33. Park, S.K., Kim, J., Kim, J. M, Yeon, J.Y., Shim, W.S., Lee, D.W., 2015. Effects of Oral Prednisolone on Recovery After Tonsillectomy. The Laryngoscope. 125, 111 – 117.

34. Perrott, D.A., Goodenough, B., Champion, G.D., 2004. Children's ratings of the intensity and unpleasantness of post-operative pain using facial expression scales. Eur J Pain. 8, 119–127.

35. Pesudovs, K., Noble, B.A., 2005. Improving Subjective Scaling of Pain Using Rasch Analysis. J Pain. 6, 630-636.

36. Powell, C.V., Kelly, A.M., Williams, A., 2001. Determining the minimum clinically significant difference in visual analog pain score for children. Ann Emerg Med. 37, 28-31.

37. Reise, S.P., Waller, N.G., 2009. Item Response Theory and Clinical Measurement. Annu Rev Clin Psychol. 5, 27–48.

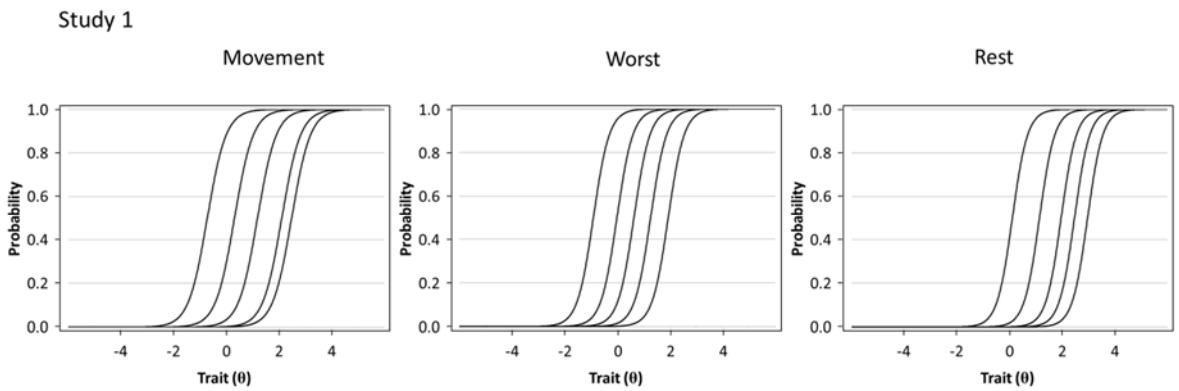
38. Samejima, F., 1969. Estimation of ability using a response pattern of graded scores. Psychometrika. 34, Supplement 1, 1–97

39. Sánchez-Rodríguez, E., Miró, J., Castarlenas, E., 2012. A comparison of four self-report scales of pain intensity in 6- to 8-year-old children. Pain. 153, 1715–1719.

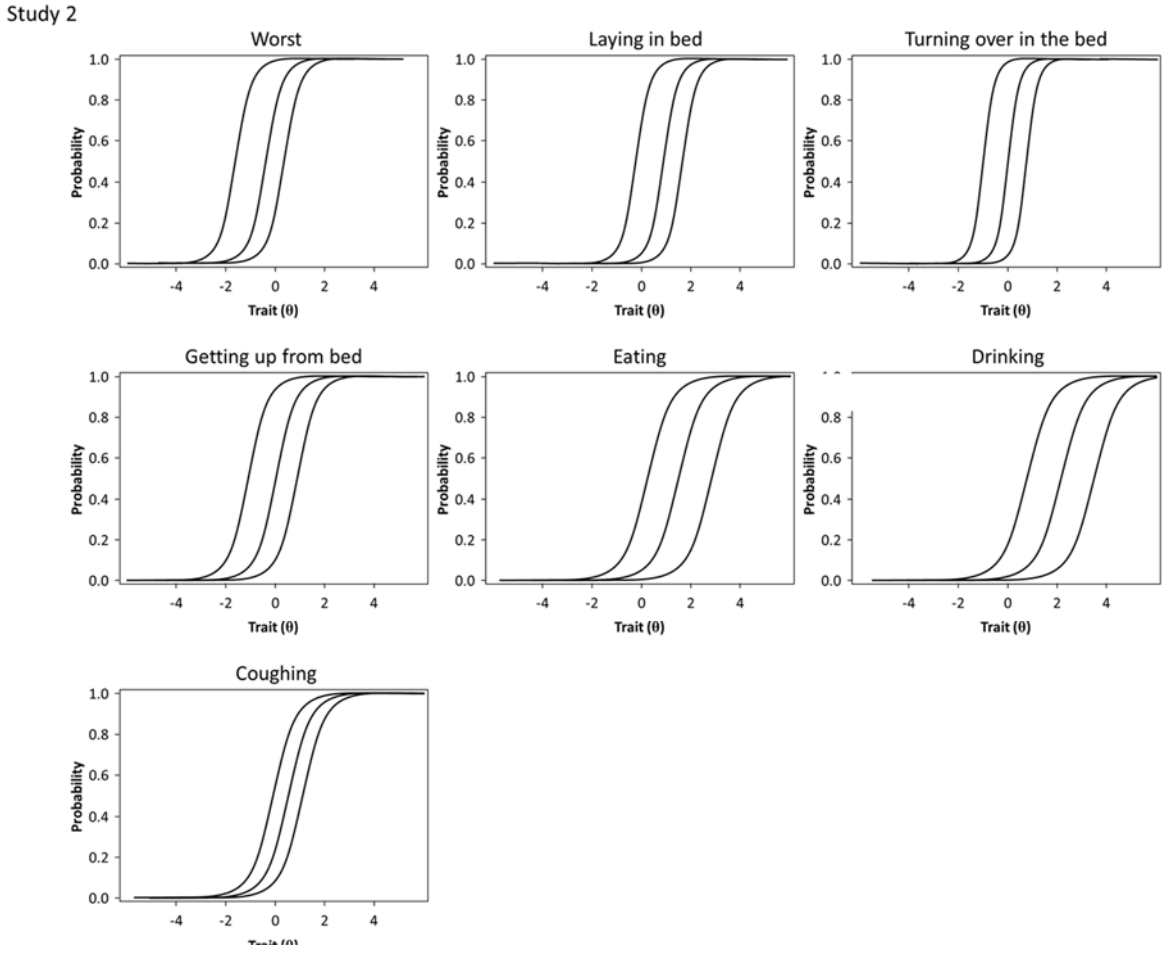
40. Shields, B.J., Cohen, D.M., Harbeck-Weber, C., Powers, J.D., Smith, G.A., 2003. Pediatric Pain Measurement Using a Visual Analogue Scale: A Comparison of Two Teaching Methods. Clin Pediatr. 42, 227-234.
41. Shields, B.J., Palermo, T.M., Powers, J.D., Grewe, S.D., Smith, G.A., 2003. Predictors of a child's ability to use a visual analogue scale. Child Care Health Dev. 29, 281–290.
42. Sijtsma, K., 2004. Item response theory, in Lewis-Beck, M., Bryman, A. E., Liao, T. F. (Eds.), The Sage encyclopedia in social science research methods. Sage Publications, Thousand Oaks CA, pp. 529-533.
43. Stanford, E.A., Chambers, C.T., Craig, K.D., 2006. The role of developmental factors in predicting young children's use of a self-report scale for pain. Pain. 120, 6–23.
44. Stoica, P., Selén, Y., 2004. Model-order selection: a review of information criterion rules. IEEE Signal processing magazine. 21, 36-47.
45. Tsze, D.S., von Baeyer, C.L., Bulloch, B., Dayan, P.S., 2013. Validation of Self-Report Pain Scales in Children. Pediatrics. 132, e971–e979.
46. von Baeyer, C.L., 2009. Children's self-report of pain intensity: What we know, where we are headed. Pain Res Manag. 14, 39–45.
47. Waterman, C., Victor, T.W., Jensen, M.P., Gould, E.M., Gammaitoni, A.R., Galer, B.S., 2010. The Assessment of Pain Quality: An Item Response Theory Analysis. The Journal of Pain. 11, 273-279.

Supplemental Figures 1: Category boundary curves for (a) study 1, (b) study 2, and (c) study 3.

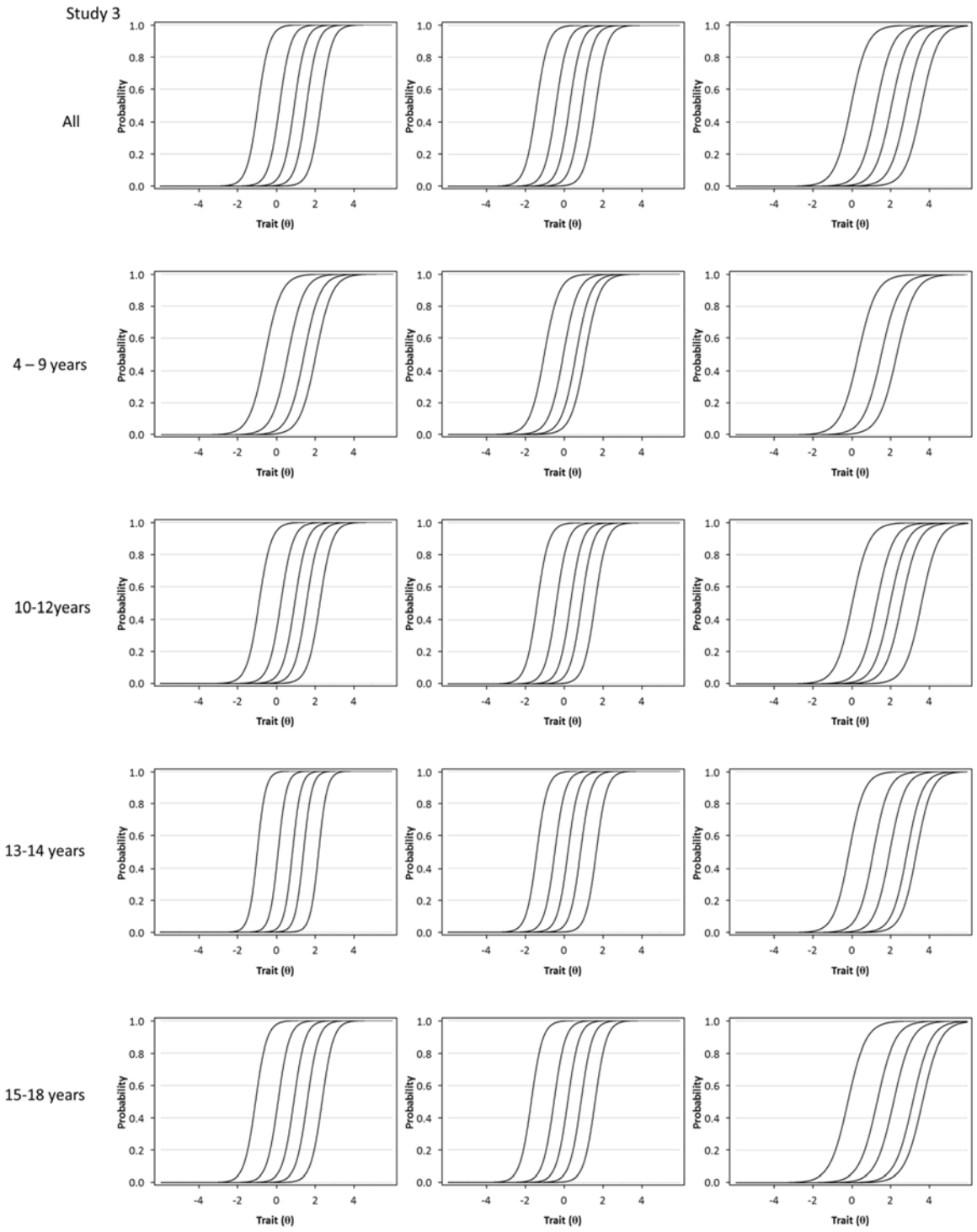
Supplemental Figures 1a:



Supplemental Figures 1b:



Supplemental Figures 1c:



Supplemental Figures 2: Category response curve for subgroups in study 3.