Published in final edited form as:

Qual Life Res. 2016 January ; 25(1): 13–23. doi:10.1007/s11136-015-1058-8.

Estimating Minimally Important Difference (MID) in PROMIS Pediatric Measures using the Scale-Judgment Method

David Thissen^a, Yang Liu^a, Brooke Magnus^a, Hally Quinn^a, Debbie S. Gipson^b, Carlton Dampier^c, I-Chan Huang^{d,1}, Pamela S. Hinds^e, David T. Selewski^b, Bryce B. Reeve^f, Heather E. Gross^g, and Darren A. DeWalt^h

David Thissen: dthissen@email.unc.edu; Yang Liu: liuy0811@live.unc.edu; Brooke Magnus: brooke.magnus@unc.edu; Hally Quinn: hallyq@live.unc.edu; Debbie S. Gipson: dgipson@med.umich.edu; Carlton Dampier: cdampie@emory.edu; I-Chan Huang: i-chan.huang@stjude.org; Pamela S. Hinds: PSHinds@childrensnational.org; David T. Selewski: dselewsk@med.umich.edu; Bryce B. Reeve: bbreeve@email.unc.edu; Heather E. Gross: hgross@email.unc.edu; Darren A. DeWalt: dewaltd@med.unc.edu

^aDepartment of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

^bDivision of Nephrology, Department of Pediatrics and Communicable Diseases, C.S. Mott Children's Hospital, University of Michigan, Ann Arbor, MI, United States

^cDepartment of Pediatrics, Emory University College of Medicine, Atlanta, GA, United States

^dInstitute for Child Health Policy, University of Florida, Gainesville, FL, United States

^eChildren's National Health System, Washington, DC and The George Washington University, Washington, DC, United States

^fDepartment of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

^gCecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

^hDivision of General Medicine and Clinical Epidemiology, Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Abstract

Objective—To assess minimally important differences (MID) for several pediatric self-report item banks from the National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System[®] (PROMIS[®]).

Methods—We presented vignettes comprising sets of two completed PROMIS questionnaires and asked judges to declare whether the individual completing those questionnaires had an important change or not. We enrolled judges (including adolescents, parents, and clinicians) who

Correspondence concerning this article should be addressed to: David Thissen, Department of Psychology, University of North Carolina at Chapel Hill, 358 Davie Hall, CB #3270, Chapel Hill, NC 27599. Telephone: 919-962-5036. dthissen@email.unc.edu.

¹I-Chan Huang's present affiliation and address is Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN, United States

Disclosure: Darren DeWalt has copyright of the items in the PROMIS scales tested in this study. He has granted unrestricted license for use of the items to the PROMIS Health Organization and receives no payment or royalties for their use.

responded to 24 vignettes (six for each domain of depression, pain interference, fatigue, and mobility). We used item response theory (IRT) to model responses to the vignettes across different judges and estimated MID as the point at which 50% of the judges would declare an important change.

Results—We enrolled 246 judges (78 adolescents, 85 parents, and 83 clinicians). The MID estimated with clinician data was about 2 points on the PROMIS *T*-score scale, and the MID estimated with adolescent and parent data was about 3 points on that same scale.

Conclusions—The MIDs enhance the value of PROMIS Pediatric measures in clinical research studies to identify meaningful changes in health status over time.

Keywords

PROMIS; pediatrics; self-report; patient-reported outcomes; item response theory; minimally important difference

Introduction

A *minimally important difference* (MID) is defined as the “smallest difference in score ... that patients perceive as important, ... and which would lead the clinician to consider a change in the patient’s management” [1–3]. MIDs are important reference values that are used to evaluate the effectiveness of interventions in clinical research. Recent recommendations for determining MIDs for patient-reported outcome (PRO) measures have emphasized two classes of procedures: distribution-based methods and anchor-based methods [2].

As Revicki et al. (p. 106) observed, “The *distribution-based indices* provide no direct information about the MID. They are simply a way of expressing the observed change in a standardized metric” [2]. This is not to say that distribution-based computations are irrelevant; but they primarily provide a source of information about the reasonableness of judgmentally determined MID values. Distribution-based methods do not identify any particular value for the MID, but differences in standard deviation units smaller than 0.2 are not likely *important*, and differences larger than 0.5 are not likely *minimal*.

Anchor-based methods use a clinical test or expert or patient judgment to divide respondents into two or more clinically meaningful categories—most straightforwardly, those that have changed and those that have not. Then, the MID is the change score on the PRO measure for those who have minimally changed on the anchor [2]. Anchor-based methods provide the current “gold standard” for the determination of MID. However, they have the disadvantage that anchor data must be available; for some domains, physiological measures serve that function, but for others, such as emotional distress, pain, or fatigue, the questionnaire measures themselves provide the primary data. Moreover, the anchors are rarely gold standards of differences that are important to patients.

Other methods have been used to select MID values; as many as nine procedures have been cataloged [4]. Many of these methods make use of expert judges, as in applications of the Delphi method to obtain consensus about the value of MID [5–9]; more recently the Delphi

method has been used as an adjunct to an anchor-based method to combine disparate values [4; 10]. Surveys of physicians have also been used to determine MID values without any attempt to build consensus [11]. Expert panels have been asked to indicate a change that is a MID on visual analog scales, or by selection of changes to item responses [12–14].

This study introduces a novel method called the *scale-judgment method* to estimate a MID for several pediatric self-report measures from the National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System® (PROMIS®). The scale-judgment method is loosely modeled after the *body of work method* in the educational measurement literature [15] and is related to several earlier expert-judgment methods [12–14]. In the scale-judgment method, panels of judges evaluate pairs of completed PRO questionnaires (generically labeled “before” and “after”); each judge indicates whether the amount of change indicated by the responses on the “before” and “after” questionnaires represents an “important” difference. Each stimulus pair also has a difference score; those values are not revealed to the judges.

The PROMIS scales are based on item response theory (IRT) calibrations, so we can generate vignettes that are plausible pairs of completed questionnaires. We do this by selecting a pair of levels for the PRO measures (e.g., pretest at 1.5 standard units above the mean, posttest at 1.0). Then we use the fact that the IRT model gives the probability for each response pattern at any level of the latent variable to create a pair of completed questionnaires that are likely, and associated with scores near the selected pairs of values. The resulting pairs of questionnaires appear to be from a longitudinal study; however, we can do these activities without any clinical data. Cella and colleagues used similarly completed individual questionnaires to collect expert judgment data to identify categories of severity based on PRO scores [16; 17].

This procedure differs from an anchor-based procedure in two ways: (1) as an advantage, these data are much easier to collect. Each judge can evaluate a large number of pairs quickly. There is no waiting for patients to change health status. (2) As a disadvantage, judges have limited information for their categorical judgment: They have only the responses to the PRO measure. The scale-judgment method differs from the use of the Delphi method in that larger groups of judges are used for scale judgment, and statistical averaging replaces any attempt to build consensus.

For pediatric measures, there are three potential groups of judges: (1) clinicians who regularly see patients; (2) parents or caregivers; and (3) adolescent patients. Each of these groups has a different perspective. In this study, we use the scale-judgment method to estimate the MID for four PROMIS pediatric measures and examine how that estimate differs across groups of judges.

Method

Research Participants

Clinicians, parents, and adolescents were recruited to evaluate change in patients' health status by looking at questionnaire responses. Research participants were recruited at four

clinical sites that had previous experience using PROMIS measures in pediatric populations with specific diseases: Children's National Health System (cancer), Emory University (sickle cell disease), University of Florida (asthma), and University of Michigan (nephrotic syndrome). University of Michigan collaborated with clinics at Duke University and Levine Children's Hospital in Charlotte, NC to enroll adolescents with nephrotic syndrome and parents of such children.

At Children's National, Emory, and the University of Michigan, researchers approached parents and adolescents in clinics. The researchers explained the study, and if parents and adolescents were eligible and willing to participate, they were provided with the link for the survey website. At the University of Florida, parents and adolescents who had participated in an earlier PROMIS study were called on the telephone to discuss this study. Parents and adolescents received gift cards valued at \$10 to \$20 for their participation.

Clinicians were identified using a variety of techniques: attendees at professional meetings; listings in physician dictionaries from the same university and state as the investigator; and colleagues of investigators in the same specialty field and/or hospital affiliation. Recruitment was done through in-person contacts, emails, and letters. Clinicians received gift cards valued at \$25 to \$50 for their participation. The study received IRB approval from regulatory boards at participating institutions.

Vignettes and Data Collection

Six pairs of completed questionnaires were created using the recommended eight- or ten-item short forms for each of four domains measured by the PROMIS pediatric scales: Depressive Symptoms [18], Pain Interference [19], Fatigue [20], and Mobility [21]. The left questionnaire of each pair was labeled "One month ago" and the right was labeled "Today." Within each set of six pairs, three were associated with improving scores, and three with worsening scores. The pairs of questionnaires were completed with responses that met several constraints: (a) the responses yielded scores that were approximately 2.5, 5, and 7.5 points different on the *T*-score scale ($M = 50$, $SD = 10$) used for the PROMIS instruments (with variation due to meeting other constraints); (b) the scores for both questionnaires were within the central part of the distribution (i.e., between 30 and 80 on the *T*-score scale); (c) the responses were likely for that score; and (d) responses to every item on each scale either changed in the same direction as the total score, or did not change.

Figure 1 shows an example of one of the 24 stimulus pairs; this pair indicates an improvement of 3.2 points. In the particular example in Figure 1, the left questionnaire has the same response for all 8 items; that pattern was not common, but it did happen in some vignettes.

The instructions were developed iteratively, with advice and feedback from researchers at all of the sites. At preliminary stages we also had informal feedback from non-clinicians, and data obtained from a small pilot study involving teens, parents, and clinicians suggested further changes. In the end, we chose not to use the word "important" in the instructions, because that seemed to have different meanings for different people. The final instructions emphasized differences that are just large enough to be noticeable. We also included a

sentence (“Looking for patterns (e.g., locations of red circles) may be misleading.”) to discourage respondents from simply declaring any difference in responses at all to be “different” as opposed to “exactly the same.”

In the main study, overall instructions read, “This is a study looking at a child’s responses to a survey about how they are doing and feeling. For each questionnaire, you will see a set of responses from 1 month ago and a set of responses from today. The items on the questionnaires from one month ago and from today are exactly the same, although the child’s responses may be different. We would like you to decide whether you think the child is doing or feeling better, worse, or about the same.” The participants were then shown an example of a pair of questionnaires, followed by more instructions: “Please read the questions and responses carefully. Looking for patterns (e.g., locations of red circles) may be misleading. Then, decide if the child is doing at least a little better today, essentially no different, or at least a little worse today.”

The stimulus pairs for each domain were preceded by an introduction for that domain: “The next set of ratings is about the child’s *<the domain name>*.” This statement was followed by an explanation of the direction that indicated worse functioning in that domain, accompanied by a graphic with the better and worse responses labeled. Then, each of the 24 stimulus pairs was presented in a frame with the instructions, “Below are the child’s responses to questions about *<the domain name>* one month ago and today.” Below the stimulus pair, the respondent was given instructions to “Please decide if you think these responses show that this child is...” followed by a list of the three response alternatives (“at least a little better today”, “essentially no different”, and “at least a little worse today”) with adjacent buttons for participants to select their response.

The survey was administered online using the Qualtrics Survey software, (<http://www.qualtrics.com>) [22]. All respondents were administered demographic (gender, age, Hispanic ethnicity, race, and primary language) and other items. Participants were then given the instructions for the MID task. The domains (e.g., depression, fatigue, pain interference, and mobility) were presented in random order, and the individual stimulus pairs were administered randomly within the domains.

Statistical Methods

The data were examined to check that there was variation in the responses for all 24 stimulus pairs, and that no data were out of bounds. Mindful of the possibility that some responders may engage in mischievous responding [23], the response patterns across the 24 stimulus pairs were examined to detect individuals who were probably not taking the judgment task seriously. While no examination of response patterns can detect all forms of mischievous or thoughtless responding, some patterns are obvious (e.g., responding “at least a little better today” for all 24 pairs). Given that the vignettes randomly switched between pairs “better today” and “worse today”, we set aside respondents with strings of ten or more of the same response.

Because this is the first attempt to use the scale-judgment method to estimate MID, a crucial aspect of the data was unknown *a priori*: We did not know whether (a) the judgments from

different participants could be treated as a homogeneous collection of responses, or (b) individual differences in the judges' criteria for "different" might have consequences for the data. In the latter case, some persons may have a consistently higher criterion than others for responding that two protocols differ. There was no way to decide *a priori* whether individual differences in the criterion would be an important feature of the data, so the data analysis proceeded along two parallel tracks, one for the homogenous assumption (a) and the other for individual differences (b).

Possibility (a) suggests logistic regression of the difference judgments on the change score values. The parameters of the logistic regression model could be used to estimate the scale score difference between vignettes corresponding to a probability of 0.5 that the pair is judged "different"; that would be the MID value. The basic model would be fitted to the data for all domains together; subsequent analyses would add dummy variables to test the differences among the domains, and between positive and negative changes. The upper panel of Figure 2 shows a hypothetical plot of the probability. A pair is judged "different" as a logistic regression function of the scale score difference associated with the pair, with a dashed line indicating the MID (~2.4) at $P = 0.5$.

To decide whether to analyze the data entertaining possibility (b), the techniques of traditional test theory and IRT apply. We compute internal consistency reliability (coefficient) of the 24-item "test" on which the items are the judged questionnaire pairs with each item having 2 response categories (different, no different). If there are negligible individual differences, then coefficient will be near zero, and the analysis can proceed as per possibility (a). If the 24 stimulus pairs taken as a test yields a reliable measure of individual difference variation, then such individual differences are an important aspect of the data, and analysis proceeds per possibility (b).

If individual differences exist in the propensity to respond "different" to a pair of completed questionnaires, then IRT can be used to produce an estimate of MID, using a definition of MID that it is the scale score difference for which an average respondent would have a 50–50 chance of responding "different". The lower panel of Figure 2 illustrates this: On the x -axis is the latent variable, which refers to the "propensity to respond 'different'"; the y -axis is the probability of responding "different." The lower panel of Figure 2 shows IRT trace lines for six stimulus pairs, associated with scale score differences between 1.1 and 5.7 points in absolute value. The dotted line shows a normal population distribution for variation in the "propensity to respond 'different'." The vertical dashed line is at the mean, and an inferred trace line is drawn as a thicker line with its location (b) parameter equal to the mean. The (hypothetical) value of the scale score difference associated with the 50–50 point of that thicker trace line, computed by interpolation of the relationship between the b parameters and scale score differences, is the estimate of the MID. In statistical analyses to identify the scale score difference associated with a b parameter equal to each population mean, the relation between the scale score differences and the b parameters would be smoothed with polynomial regression. As with the logistic regression analysis described above, terms for domains and positive vs. negative changes could be tested for significance in this regression analysis.

Results

Table 1 summarizes demographic characteristics of the respondents.

Data for 19 of the 246 respondents were set aside because they appeared to be mischievous or less than thoughtful responses. Ten respondents set aside were adolescents and nine were parents; none were clinicians. Subsequent analyses are based on data from the remaining 227 respondents.

Table 2 lists the 24 stimulus pairs, with the overall frequencies that responded “At least a little better today” (hereafter “better”), “Essentially no different” (hereafter “no difference”) and “At least a little worse today” (hereafter “worse”). A substantial proportion of the responses (4–20%) are in the “wrong direction.” For Depressive Symptoms, Pain, and Fatigue, a positive difference is worse; scores on Mobility are better if the difference is positive. For each of the stimulus pairs, the “today” responses are unambiguously either “better” or “worse.” So all of the responses should be better-or-no-different, or worse-or-no-different, for any given stimulus pair. The column of Table 2 labeled “proportion wrong direction” shows that proportion for each pair.

There are two possible explanations for the wrong-direction responses: One is that some judges were careless; this explanation suggests that the responses in the wrong direction should be omitted from subsequent analyses. A second explanation is that, because scoring direction varies, some judges were sometimes confused, and responded “better” when they meant “worse,” or vice versa. That explanation suggests data analysis with the wrong-direction responses reversed to mean different (in the correct direction). To cover both possibilities we repeat subsequent analyses, with wrong-direction responses omitted and reversed.

We computed internal consistency reliability of the 24-item “test” on which the items are the questionnaire pairs; coefficient is 0.81 if “wrong direction” responses are omitted, and 0.83 if they are reversed. So the 24-vignette form is a moderately reliable test of individual differences in the propensity to say that two sets of response patterns differ, and we proceed with the individual differences analysis.

We fitted the 1-parameter logistic (1PL) IRT model to the data, considering the set of stimulus pairs a 24-item test of the propensity to say response patterns differ. The data were divided into three populations: clinicians, adolescents, and parents. In the formulation of the 1PL model [24] used here, one common slope for all items (stimulus pairs) was estimated, along with the means and standard deviations of normal distributions of the latent variable for the adolescent and parent groups relative to the scale-defining mean of 0 and standard deviation of 1 for the clinicians. For the analysis with the “wrong direction” responses reversed, the slope parameter estimate is 1.33 (*s.e.* = 0.26), and the means and standard deviations are = −0.36 and = 1.51 for the adolescent group and = −0.53 and = 1.47 for the parents. For the analysis with the “wrong direction” responses omitted, the slope parameter estimate is 1.17 (*s.e.* = 0.29), and the means and standard deviations are = −0.18 and = 1.42 for the adolescent group and = −0.37 and = 1.31 for the parents. These values indicate that individual differences exist among the respondents in the propensity to say that two sets of

response patterns differ (the slope estimate would be 0.0 otherwise), and that the adolescents and parents are, on average, lower on that latent variable than the clinicians, and more variable. Likelihood-ratio tests of the differences among the three groups' means and standard deviations are significant for the wrong-direction-omitted analysis ($G^2=15.85$, 4 *d.f.*, $p = 0.003$) and nearly so for the wrong-direction-reversed analysis ($G^2=8.82$, 4 *d.f.*, $p = 0.065$).

The lower panel of Figure 2 shows the results of a subset of these IRT analyses. In the graphic, the reference distribution is for the clinicians, and the six trace lines are for a subset of 6 of the 24 stimulus pairs, obtained with the “wrong direction” responses reversed. The b parameters (thresholds) for all 24 stimulus pairs are in Table 3, for both the wrong-direction-omitted and wrong-direction-reversed analyses. Figure 3 shows the absolute value of the difference between the scale scores for each stimulus pair plotted against those b parameters. Note that the rank order of the absolute differences generally coincides with the rank order of b parameters. Smooth curves have been added to those plots using quadratic regression of the scale score difference on the b values. We also considered regressions including indicator variables for the domains, and for positive vs. negative change; none of the latter coefficients differed significantly from zero, so we used one curve to smooth the data for all four domains, and for positive and negative change.

We used the idea illustrated in the lower panel of Figure 2 to compute estimates of MID, with wrong direction responses omitted, and with wrong direction responses reversed. The goal is to estimate the scale score difference for a (hypothetical) pair of filled-in questionnaires that would have a b parameter of equal to the mean of the respondents' distribution; that is, that an average respondent would say “different” with a probability of 50%. Because the intercept terms in the quadratic regression models illustrated in Figure 3 represent the scale score difference at 0.0, those intercepts are the desired MID estimates for clinicians, and their locations are shown with dashed lines. For wrong direction responses omitted, clinicians' MID = 2.11, s.e. 0.59, adolescents' MID = 2.25, s.e. 0.61, and parents' MID = 2.38, s.e. 0.66; and for wrong direction responses reversed, clinicians' MID = 1.91, s.e. 0.56, adolescents' MID = 2.12, s.e. 0.61, and parents' MID = 2.19 s.e. 0.68.

Discussion

The PROMIS pediatric measure's MID is about 2 points, with a standard error of a little over a half point, for the average clinician. For the average adolescent, or the average parent, MID is a little higher. A sensitivity analysis treating the “wrong directions responses” differently (omitting them versus reversing their direction) yielded similar MID estimates. These values are near the MIDs between 2.4 and 3.5 points for longitudinal anchor-based analyses reported by Yost et al. [25] for a collection of the adult PROMIS measures studied with cancer patients and survivors.

The instructions to the research participants focused on noticeable changes, so in this study the definition of a MID is very similar to a just-noticeable difference. Given the heterogeneity of the judges, it is not clear how the instructions could be modified to reliably instruct all participants to somehow distinguish between “just noticeable” and “minimally

important”. The instructions resolve that issue by focusing on the just-noticeable difference, in the spirit of the recommendation that MID be defined as minimum detectable change to avoid proliferation of different MID values from different procedures [26; 27].

A significant limitation of the scale-judgment method is the same as its primary advantage with respect to anchor-based methods: The scale-judgment method uses judgments about hypothetical situations; there is no reference to the actual change of real persons. If it is practical to obtain valid anchoring data, an anchor-based method is almost certainly preferred. However, it may be desirable to obtain an estimate of a MID before anchoring data can be obtained, or for scales for which there is no clear anchor.

Because this was the first empirical use of the scale-judgment method, we learned a great deal in this study that can be applied to improve future use of the method. First, we found that individual difference variation among the judges is an important source of variability in the data, so we made use of IRT to model those individual differences, and defined MID as the value of the scale score difference for which the probability of a “different” judgment is 0.5. IRT is used twice in this procedure: once to construct the original scales, and a second time to analyze the scale-judgment data.

Second, we learned that mixing scales for which a higher score is an improvement with others for which a lower score is an improvement in the same judgment task is not recommended. Mixing led to the possibility that judges might misremember which direction represents “better” or “worse” for a particular scale. In future applications of this method, we would collect data on only one scale at a time, which would naturally be the case if this were a routine part of scale development. Then judges would not have the opportunity to confuse the direction of improvement within the judgment task.

Using only six questionnaire pairs for each domain, we were unable to detect statistically significant differences among MID values across domains. Such differences may exist. In order to measure such differences, more stimulus pairs would be needed to obtain sufficiently precise estimates of the MID values for each domain.

We showed that the value of MID differs across groups of judges: MID is about 2 points for the average clinician, but higher for adolescents and parents. Although this has not been widely studied, other investigators have found similar results [17; 28]. These findings suggest that patients may expect greater change before taking action.

Acknowledgments

PROMIS® was funded with cooperative agreements from the National Institutes of Health (NIH) Common Fund Initiative (Northwestern University, PI: David Cella, PhD, U54AR057951, U01AR052177; Northwestern University, PI: Richard C. Gershon, PhD, U54AR057943; American Institutes for Research, PI: Susan (San) D. Keller, PhD, U54AR057926; State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, U01AR057948, U01AR052170; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, U01AR057954; University of Washington, Seattle, PI: Dagmar Amtmann, PhD, U01AR052171; University of North Carolina, Chapel Hill, PI: Harry A. Guess, MD, PhD (deceased), Darren A. DeWalt, MD, MPH, U01AR052181; Children’s Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, U01AR057956; Stanford University, PI: James F. Fries, MD, U01AR052158; Boston University, PIs: Alan Jette, PT, PhD, Stephen M. Haley, PhD (deceased), and David Scott Tulskey, PhD (University of Michigan, Ann Arbor), U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD (University of Michigan, Ann Arbor) and Brennan Spiegel, MD, MSHS, U01AR057936; University of

Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR052155; Georgetown University, PIs: Carol M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle) and Arnold L. Potosky, PhD, U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan DeWitt, MD, MSCE, U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, U01AR052186). NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD (deceased), Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Peter Scheidt, MD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztejn, MD, William Phillip Tonkins, DrPH, Ellen Werner, PhD, Tisha Wiley, PhD, and James Witter, MD, PhD. We thank Catriona Mowbray, PhD, RN at the Children's National Health System site, as well as Susan Massengill at Levine Children's Hospital and Rasheed Gbadegesin at Duke University for important contributions, and Karon Cook, PhD, and Ron Hays, PhD, for very helpful comments on an earlier draft. The contents of this article uses data developed under PROMIS. These contents do not necessarily represent an endorsement by the US Federal Government or PROMIS. See www.nihpromis.org for additional information on the PROMIS® initiative.

We would also like to acknowledge Susan Massengill at Levine Children's Hospital and Rasheed Gbadegesin at Duke University, as they led the MID work for sampling the nephrotic syndrome population at these institutions.

Abbreviations

PROMIS®	Patient-Reported Outcomes Measurement Information System®
NIH	National Institutes of Health
MID	Minimally important difference
PRO	Patient-Reported Outcome

References

1. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc.* 2002; 77(4):371–383.10.4065/77.4.371 [PubMed: 11936935]
2. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol.* 2008; 61(2):102–109.10.1016/j.jclinepi.2007.03.012 [PubMed: 18177782]
3. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989; 10(4):407–415.10.1016/0197-2456(89)90005-6 [PubMed: 2691207]
4. Wells G, Li T, Maxwell L, MacLean R, Tugwell P. Determining the minimal clinically important differences in activity, fatigue, and sleep quality in patients with rheumatoid arthritis. *J Rheumatol.* 2007; 34(2):280–289. [PubMed: 17304654]
5. Bellamy N, Anastassiades TP, Buchanan WW, Davis P, Lee P, McCain GA, Wells GA, Campbell J. Rheumatoid arthritis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs--results of a consensus development (Delphi) exercise. *J Rheumatol.* 1991; 18(12):1908–1915. [PubMed: 1795330]
6. Bellamy N, Buchanan WW, Esdaile JM, Fam AG, Kean WF, Thompson JM, Wells GA, Campbell J. Ankylosing spondylitis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs--results of a consensus development (Delphi) exercise. *J Rheumatol.* 1991; 18(11):1716–1722. [PubMed: 1787494]
7. Bellamy N, Carette S, Ford PM, Kean WF, le Riche NG, Lussier A, Wells GA, Campbell J. Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials--results of a consensus development (Delphi) exercise. *J Rheumatol.* 1992; 19(3):451–457. [PubMed: 1578462]
8. Spiegel BM, Younossi ZM, Hays RD, Revicki D, Robbins S, Kanwal F. Impact of hepatitis C on health related quality of life: a systematic review and quantitative assessment. *Hepatology.* 2005; 41(4):790–800.10.1002/hep.20659 [PubMed: 15791608]

9. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. *Health Serv Res.* 2007; 42(6 Pt 1):2257–2274. discussion 2294–2323. 10.1111/j.1475-6773.2007.00733.x [PubMed: 17995565]
10. Rai SK, Yazdany J, Fortin PR, Avina-Zubieta JA. Approaches for estimating minimal clinically important differences in systemic lupus erythematosus. *Arthritis Res Ther.* 2015; 17:143.10.1186/s13075-015-0658-6 [PubMed: 26036334]
11. van Walraven C, Mahon JL, Moher D, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *J Clin Epidemiol.* 1999; 52(8):717–723.10.1016/S0895-4356(99)00050-5 [PubMed: 10465315]
12. Todd KH, Funk JP. The minimum clinically important difference in physician-assigned visual analog pain scores. *Acad Emerg Med.* 1996; 3(2):142–146.10.1111/j.1553-2712.1996.tb03402.x [PubMed: 8808375]
13. Dempster H, Porepa M, Young N, Feldman BM. The clinical meaning of functional outcome scores in children with juvenile arthritis. *Arthritis Rheum.* 2001; 44(8):1768–1774.10.1002/1529-0131(200108)44:8<1768::AID-ART312>3.0.CO;2-Q [PubMed: 11508427]
14. Gong GW, Young NL, Dempster H, Porepa M, Feldman BM. The Quality of My Life questionnaire: the minimal clinically important difference for pediatric rheumatology patients. *J Rheumatol.* 2007; 34(3):581–587. [PubMed: 17183614]
15. Kingston, NM.; Kahl, SR.; Sweeney, KP.; Bay, L. Setting performance standards using the body of work method. In: Cizek, GJ., editor. *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2001. p. 219-248.
16. Cella D, Choi S, Garcia S, Cook KF, Rosenbloom S, Lai JS, Tatum DS, Gershon R. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Qual Life Res.* 2014;10.1007/s11136-014-0732-6
17. Cook KF, Victorson DE, Cella D, Schalet BD, Miller D. Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning and sleep disturbance using standard setting with patients and providers. *Qual Life Res.* 2015; 24(3):575–589.10.1007/s11136-014-0790-9 [PubMed: 25148759]
18. Irwin DE, Stucky B, Langer MM, Thissen D, Dewitt EM, Lai JS, Varni JW, Yeatts K, DeWalt DA. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research: an international journal of quality of life aspects of treatment, care and rehabilitation.* 2010; 19(4):595–607.10.1007/s11136-010-9619-3
19. Varni JW, Stucky BD, Thissen D, Dewitt EM, Irwin DE, Lai JS, Yeatts K, Dewalt DA. PROMIS Pediatric Pain Interference Scale: An item response theory analysis of the Pediatric Pain Item Bank. *Journal of Pain.* 2010; 11(11):1109–1119.10.1016/j.jpain.2010.02.005 [PubMed: 20627819]
20. Lai JS, Stucky BD, Thissen D, Varni JW, Dewitt EM, Irwin DE, Yeatts KB, Dewalt DA. Development and psychometric properties of the PROMIS(R) pediatric fatigue item banks. *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation.* 2013;10.1007/s11136-013-0357-1
21. Dewitt EM, Stucky BD, Thissen D, Irwin DE, Langer M, Varni JW, Lai JS, Yeatts KB, Dewalt DA. Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: Built using item response theory. *Journal of Clinical Epidemiology.* 2011; 64(7):794–804.10.1016/j.jclinepi.2010.10.012 [PubMed: 21292444]
22. Qualtrics. *Qualtrics Research Suite (Version 37892).* Provo, UT: 2013.
23. Robinson-Cimpian JP. Inaccurate Estimation of Disparities Due to Mischievous Responders: Several Suggestions to Assess Conclusions. *Educational Researcher.* 2014; 43(4):171–185.10.3102/0013189x14534297
24. Thissen D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika.* 1982; 47(2):175–186.10.1007/bf02296273
25. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology.* 2011; 64(5):507–516.10.1016/j.jclinepi.2010.11.018 [PubMed: 21447427]

26. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J.* 2008; 8(6):968–974.10.1016/j.spinee.2007.11.006 [PubMed: 18201937]
27. Copay AG. Commentary: the proliferation of minimum clinically important differences. *Spine J.* 2012; 12(12):1129–1131.10.1016/j.spinee.2012.11.022 [PubMed: 23312139]
28. Demyttenaere K, Desai D, Petit C, Croenlein J, Brecht S. Patient-assessed versus physician-assessed disease severity and outcome in patients with nonspecific pain associated with major depressive disorder. *Prim Care Companion J Clin Psychiatry.* 2009; 11(1):8–15.10.4088/PCC.08m00670 [PubMed: 19333404]

One month ago					Today				
I felt alone.					I felt alone.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I felt like I couldn't do anything right.					I felt like I couldn't do anything right.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I felt everything in my life went wrong.					I felt everything in my life went wrong.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I felt sad.					I felt sad.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I thought that my life was bad.					I thought that my life was bad.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I could not stop feeling sad.					I could not stop feeling sad.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I felt lonely.					I felt lonely.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always
I felt unhappy.					I felt unhappy.				
never	almost never	sometimes	often	almost always	never	almost never	sometimes	often	almost always

Figure 1.
One of the 24 stimulus pairs, for the Depressive Symptoms scale; the left questionnaire is associated with a score of 62.1 and the right questionnaire has a score of 58.9, so this pair indicates improvement (less depression) of 3.2 points.

Figure 2.

Upper panel: A hypothetical plot of the probability a pair of questionnaire responses is judged “different” as a logistic regression function of the scale score difference associated with the pair, with a dashed line indicating the MID at $P = 0.5$. Lower panel: The x -axis is the latent variable (propensity to respond “different”); the y -axis is the probability of responding “different.” The curves are IRT trace lines for six stimulus pairs, associated with scale score differences between 1.1 and 5.7 points. The normal population distribution is shown as a dotted line; the vertical dashed line is at the mean of the reference population, and an inferred trace line is drawn as a thicker line with its location (b) parameter equal to 0.0. The scale score difference associated with that thicker trace line is the estimate of the MID.

Figure 3.

The absolute scale score difference between the two halves of each stimulus pair plotted against the b parameters for the scaled judgment stimulus pairs (wrong direction omitted in the upper panel, and wrong direction reversed in the lower panel). Smooth curves have been added to those plots using quadratic regression of the scale score difference on the b values; the dashed lines illustrate MID computation for the clinicians' average score (0.0). MID values for the adolescents and parents are computed similarly, as the value of the curve at the average adolescent values (-0.18 wrong direction omitted, -0.36 wrong direction

reversed) and the average parent values (-0.37 wrong direction omitted, -0.53 wrong direction reversed).

Note: The point for the stimulus pair “Mobility 6” has been omitted from Figure 3, and is not used in the data analysis. For that stimulus pair, the scale score difference between “1 month ago” and “today” was 8.5 points, which is very large; however, due to the ceiling effect on the Mobility scale, that was a change of only one point for a single item, “I could do sports and exercise that other kids my age could do,” from “with no trouble” to “with a little trouble.” Very few judges (33% or 37%) considered that change to be “different.” Those data were set aside because this combination of a very small response change with a very large scale-score change is not representative of the behavior of the questionnaires as a whole.

Table 1

Demographic characteristics of the study sample.

	Adolescents N = 78 (%)	Parents/Guardians N = 85 (%)	Clinicians N = 83 (%)
Mean Age (SD)	14.9 (1.5)	42.9 (7.9)	41.6 (9.2)
Age Range	13–18	25–82	28–68
Gender			
Male	47 (60.3)	16 (18.8)	23 (27.7)
Female	31 (39.7)	69 (81.2)	60 (72.3)
Race			
White	28 (35.9)	44 (51.8)	51 (61.4)
Black or African American	39 (50.0)	35 (41.2)	14 (16.9)
Asian	4 (5.1)	2 (2.4)	10 (12.0)
Other	1 (1.3)	2 (2.4)	3 (3.6)
Multiple Races	6 (7.7)	2 (2.4)	5 (6.0)
Hispanic ethnicity			
Non Hispanic	73 (93.6)	81 (95.3)	80 (96.4)
Hispanic	5 (6.4)	4 (4.7)	3 (3.6)

Table 2

The stimulus pairs, in order of scale score difference within domain, with the frequencies of “better”, “no different”, and “worse” judgments.

Stimulus Label	Scale Score			Frequency		
	1 month ago	Today	Difference	Better	No difference	Worse
Depressive Symptoms 2	49.5	57.9	8.4	23	19	185
Depressive Symptoms 3	56.7	62.1	5.4	32	18	176
Depressive Symptoms 1	43.5	45.9	2.4	15	151	61
Depressive Symptoms 5	64.3	62.1	-2.2	133	66	27
Depressive Symptoms 4	62.1	58.9	-3.2	179	33	15
Depressive Symptoms 6	73.4	66.0	-7.4	189	21	17
Pain 1	43.6	51.4	7.8	28	18	180
Pain 2	52.4	57.6	5.2	39	13	175
Pain 3	56.8	60.4	3.6	18	33	176
Pain 4	58.5	57.6	-0.9	72	146	9
Pain 6	72.9	68.0	-4.9	174	35	17
Pain 5	68.0	59.6	-8.4	191	16	19
Fatigue 2	52.2	57.8	5.6	39	21	167
Fatigue 3	56.9	60.1	3.2	45	43	139
Fatigue 1	45.9	47.0	1.1	14	171	42
Fatigue 4	60.1	57.8	-2.3	127	87	13
Fatigue 5	68.6	62.2	-6.4	194	16	17
Fatigue 6	76.4	68.6	-7.8	179	25	22
Mobility 3	40.2	46.9	6.7	187	19	21
Mobility 2	34.0	39.5	5.5	182	16	29
Mobility 1	30.5	34.0	3.5	184	23	20
Mobility 4	46.9	44.3	-2.6	26	135	66
Mobility 5	50.0	44.3	-5.7	22	53	152
Mobility 6	58.5	50.0	-8.5	14	143	70

Note: PROMIS Scale Scores are on a *T*-score metric with mean 50 and standard deviation of 10 from the calibration population (ages 8–17). Higher scores for depressive symptoms, pain, and fatigue represent worsening symptom burden and higher scores for mobility represent better functioning.

Table 3

The stimulus pairs, in order of scale score difference within domain, with 1PL b parameters for “different” judgments.

Stimulus Label	Scale Score			Wrong Direction Omitted		Wrong Direction Reversed	
	1 month ago	Today	Difference	b	s.e.	b	s.e.
Depressive Symptoms 2	49.5	57.9	8.4	-2.57	0.55	-2.85	0.73
Depressive Symptoms 3	56.7	62.1	5.4	-2.63	0.57	-2.90	0.74
Depressive Symptoms 1	43.5	45.9	2.4	0.75	0.22	0.61	0.25
Depressive Symptoms 5	64.3	62.1	-2.2	-1.08	0.29	-1.22	0.36
Depressive Symptoms 4	62.1	58.9	-3.2	-2.14	0.47	-2.20	0.57
Depressive Symptoms 6	73.4	66.0	-7.4	-2.58	0.55	-2.74	0.70
Pain 1	43.6	51.4	7.8	-2.62	0.56	-2.91	0.75
Pain 2	52.4	57.6	5.2	-3.07	0.66	-3.26	0.83
Pain 3	56.8	60.4	3.6	-2.02	0.45	-2.20	0.57
Pain 4	58.5	57.6	-0.9	0.49	0.21	0.49	0.24
Pain 6	72.9	68.0	-4.9	-2.00	0.44	-2.13	0.56
Pain 5	68.0	59.6	-8.4	-2.83	0.60	-3.04	0.78
Fatigue 2	52.2	57.8	5.6	-2.52	0.54	-2.74	0.70
Fatigue 3	56.9	60.1	3.2	-1.55	0.37	-1.86	0.50
Fatigue 1	45.9	47.0	1.1	1.27	0.28	1.11	0.33
Fatigue 4	60.1	57.8	-2.3	-0.70	0.24	-0.76	0.28
Fatigue 5	68.6	62.2	-6.4	-2.90	0.62	-3.04	0.78
Fatigue 6	76.4	68.6	-7.8	-2.34	0.51	-2.53	0.65
Mobility 3	40.2	46.9	6.7	-2.74	0.58	-2.85	0.73
Mobility 2	34.0	39.5	5.5	-2.93	0.62	-3.04	0.78
Mobility 1	30.5	34.0	3.5	-2.54	0.54	-2.63	0.68
Mobility 4	46.9	44.3	-2.6	0.56	0.21	0.25	0.22
Mobility 5	50.0	44.3	-5.7	-1.38	0.34	-1.57	0.43
Mobility 6	58.5	50.0	-8.5	0.53	0.21	0.43	0.23

Note: PROMIS Scale Scores are on a T -score metric with mean 50 and standard deviation of 10 from the calibration population (ages 8–17). Higher scores for depressive symptoms, pain, and fatigue represent worsening symptom burden and higher scores for mobility represent better functioning. b = threshold parameter of the 1-parameter logistic (1PL) IRT model.