Comparison of Raw And Regression Approaches To Capturing Change On Patient Reported Outcome Measures

David A. Andrae (andraeda@gmail.com)
Endpoint Outcomes https://orcid.org/0000-0003-2153-1508
Brandon Foster
Endpoint Outcomes
J. Devin Peipert
Northwestern University Feinberg School of Medicine

Research Article

Keywords: Patient-reported outcome, estimators, potential covariates, thresholds

Posted Date: July 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-485296/v1

Comparison of Raw and Regression Approaches to Capturing Change on Patient Reported Outcome Measures

David A. Andrae[1]

Endpoint Outcomes

Austin, TX USA

dave.andrae@endpointoutcomes.com

ORCID iD: 0000-0003-2153-1508

Brandon Foster

Endpoint Outcomes

Boston, MA USA

brandon.foster@endpointoutcomes.com

ORCID iD: 0000-0002-8934-9860

J Devin Peipert

Northwestern University Feinberg School of Medicine, Department of Medical Social Sciences

Chicago, IL

john.peipert@northwestern.edu

ORCID iD: 0000-0001-5762-7881

Abstract

Purpose

Patient-reported outcome (PRO) analyses often involve calculating raw change scores, but limitations of this approach are well documented. Regression estimators can incorporate information about measurement error and potential covariates, potentially improving change estimates. Yet, adoption of these regression-based change estimators is rare in clinical PRO research.

[1] Corresponding Author

Methods

Both simulated and PROMIS® pain interference items were used to calculate change employing three methods: raw change scores and regression estimators proposed by Lord and Novick (LN) and Cronbach and Furby (CF). In the simulated data, estimators' ability to recover true change was compared. Standard errors of measurement (SEM) and prediction (SEP) with associated 95% confidence limits were also used to identify criteria for significant improvement. These methods were then applied to real-world data from the PROMIS® study.

Results

In the simulation, both regression estimators reduced variability compared to raw change scores by almost half. Compared to CF, the LN regression better recovered true simulated differences. Analysis of the PROMIS® data showed similar themes, and change score distributions from the regression estimators showed less dispersion. Using distribution-based approaches to calculate thresholds for significant within-patient change, smaller changes could be detected using both regression estimators.

Conclusions

These results suggest that calculating change using regression estimates may result in more increased measurement sensitivity. Using these scores in lieu of raw differences can help better identify individuals who experience real underlying change in PROs in the course of a trial, and enhance the established methods for identifying thresholds for meaningful within-patient change in PROs.

Introduction

Estimating meaningful within-patient change is among the most important elements of statistical analysis to support patient reported outcomes (PROs) as endpoints in clinical trials. In their most recent guidance for clinical outcome assessments, the United States Food and Drug Administration (FDA) defines meaningful within patient change as improvement or deterioration from the patient's perspective, and serves as a way of defining clinical benefit on a PRO [1]. Improvement and deterioration are captured in terms of change in PRO scores over the course of a clinical trial. For example, on a fatigue PRO where higher scores indicate worse fatigue, deterioration is indicated by increases in scores and improvement is indicated by a reduction in scores.

In clinical trials, change in PRO scores are calculated almost exclusively using the difference in scores from a post-baseline timepoint to a baseline timepoint—i.e., the raw change. These raw change scores

are then used in conjunction with other methods to estimate meaningful within person change. [2]. To determine meaningful thresholds for within-patient change, the FDA currently recommends stratifying raw change scores on a PRO by an anchor variable. Then, distributions of change scores by anchor group categories are visualized by plotting empirical cumulative distribution function (eCDF) and probability density function (PDF) curves [1]. While raw change scores are simple to calculate and are often easy to interpret, they have several notable disadvantages [3]. Of these disadvantages, the largest drawback of raw change scores may be their high measurement error, making them unreliable, and potentially leading to misguided conclusions [4].

Due to the problems with change scores, alternative approaches are needed. Fortunately, the classic psychometric literature has several potential directions to advance estimation of change on PROs, but these have generally gone unused in health and clinical trials. Lord offered regression estimators of a true difference on a measure over two timepoints [5–7]. Regression provides a framework for discerning true change from error. In classical test theory (CTT), any score, including change scores, is comprised of a true element and error (e.g., measurement error). Approaches that can distinguish true change from error are likely superior to the primitive difference between scores at two timepoints. An additional element of more advanced approaches to estimating change involve predicting post-test scores and determining the how much the observed post-test value deviates from the prediction [3, 8]. Notably, Lord's estimator incorporates this element of deviation from the predicted post-test value, as well as another key element, the correlation between pre- and post-test. Additionally, Cronbach and Furby extended Lord's estimator by accommodating additional variables that may improve estimation [7]. Additional variables may be measured at the pre- or post-test and can be alternative, potentially "gold standard" measures of the construct. Cronbach and Furby refer to this as complete estimation. Both of these innovations address the poor reliability of raw change scores directly by incorporating additional information other than the baseline responses in to the calculation of the change.

Accounting for Individual-level changes

Traditional treatments within CTT model observed scores as a decomposition of true score and error as in Equation 1.

Equation 1.

$$X = \mathrm{T} + e$$

With errors (ε) being independent. However, to account for repeated measures of X and the subsequent measurement, Y, we need to update Equation 1 as

Equation 2.

$$X = \mathrm{T} + e_X + \varepsilon_X$$

Where ε_X represents a random effect attributable to individuals' repeated measurements [7]. The correlations between measurements can be, based on these two models, as unlinked—i.e., based on Equation 1—and linked, based on Equation 2. The unlinked correlation is represented in Equation 3

Equation 3.

$$r_{XY} = \frac{cov(\tau_X, \tau_Y)}{\sigma_X \sigma_Y}$$

And the linked version in Equation 4.

Equation 4.

$$\rho_{XY} = \frac{cov(\tau_X, \tau_Y) + cov(\varepsilon_X, \varepsilon_Y)}{\sigma_X \sigma_Y}$$

Thus, the linkage between X and Y is taken into account with ρ_{XY}.

These linked and unlinked correlations can be used to assess the reliability of a change score [7, 8]. As both X and Y potentially have both independent and dependent measurement error components, the reliability of the difference between them should take these errors into account. Equation 5 shows how this relationship is calculated for the linked case.

Equation 5.

$$\rho_{DD'} = \frac{\sigma_X^2 \rho_{XX'} + \sigma_Y^2 \rho_{YY'} - 2\sigma_X \sigma_Y \rho_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_X \sigma_Y \rho_{XY}}$$

Note that the linked case was an extension posited by Cronbach and Furby to the Lord and Novick [7, 8] presentation in which ρ_{XY} is replaced by r_{XY}.

Regression-based estimators for individual change scores

To estimate change scores while taking measurement error and individual differences into account, one approach is to treat the problem as a regression estimation. The regression estimator described by Lord and Novick [8] and extended by Cronbach and Furby [7] is defined for linked scores as:

Equation 6.

$$\widehat{D}_{LN} = \frac{1}{1-\rho_{XY}^2}\left[\frac{Y}{\sigma_Y}(\sigma_Y\rho_{YY'} - \sigma_Y r_{XY}\rho_{XY} + \sigma_X r_{XY}\rho_{XX'} - \sigma_X\rho_{XY}) - \frac{X}{\sigma_X}(\sigma_X\rho_{XX'} - \sigma_X r_{XY}\rho_{XY} + \sigma_Y r_{XY}\rho_{YY'} - \sigma_Y\rho_{XY})\right] + C$$

Or, rearranged to look more like a regression equation as [9]:

$$\widehat{D}_{LN} = -\frac{\rho_{XX'} - r_{XY}\rho_{XY} + \frac{\sigma_Y}{\sigma_X}(r_{XY}\rho_{YY'} - r_{XY})}{1-\rho_{XY}^2}X + \frac{\rho_{YY'} - r_{XY}\rho_{XY} + \frac{\sigma_X}{\sigma_Y}(r_{XY}\rho_{XX'} - r_{XY})}{1-\rho_{XY}^2}Y + C$$

Thus

$$\widehat{D}_{LN} = -\beta_X X + \beta_Y Y + C$$

Where X is a score at time 1, Y is a score at time 2, r_{XY} is the unlinked correlation between time 1 and 2, and ρ_{XY} is the linked version of that correlation; the C term is a constant that aligns the mean of the estimator to be equal to that of the raw difference. Note that Lord and Novick did not make the linked versus unlinked distinction, but it has been here. This equation is based on the idea that the measurement error is inherent in each measurement occasion, i.e., X and Y, and so some regression to the mean is expected. The expressions that include the reliabilities, correlations, and standard deviations of X and Y are ostensibly treated as regression coefficients to account for the measurement error inherent to each administration of X and Y.

A further extension of this estimator was also developed by Cronbach and Furby [7], and is described as the *complete estimator*. This estimator further builds on the prior work of Lord and Novick [8] by also adjusting for covariates, thereby incorporating validating information into the change score estimation. The idea was to take the basic notion of accounting for measurement error in the target scores and add relevant information from other measurements to form a more precise estimate of change. The estimator is described as:

Equation 7.

$$\widehat{D}_{CF} = -\beta_X X + \beta_Y Y + \beta_i W_i + \beta_j Z_j + C$$

Where the β coefficients for X and Y and C are defined as above. The other two terms represent residual scores of the partial variate of the other terms. The covariates are additional information about each respondent with W_i being i variables collected concurrently with X and Z_j similarly for Y. The covariates are entered into the estimate as partial variates, i.e., residual scores from regressing W_i and Z_j onto the other variables taken into account. Thus, the approach, to the extent relevant covariates are included, adds to the information used to estimate a change score.

Methods

The main objective of the current study is to compare the recovery of real population differences in individuals' scores across two time points using Lord and Novick's regression estimator and Cronbach and Furby's complete estimator. These two regression-based estimators for calculating individual change scores are compared against the calculation of raw differences. First, we present a simulation study as a proof of concept and to demonstrate the performance for the regression-based estimators of individual differences in a context where the true population theta is known. Next, the applicability of these methods is demonstrated using data from the PROMIS® 1 Wave 2 Depression and Pain validation study. The applicability of the methods presented herein can apply to any number of scoring algorithms use to generate scores cross-sectionally and applied in the analysis of change overtime. More technical details for these methods can be provided upon request.

Simulation study

Simulated item response data were generated for 200 respondents. A total of 21 items were simulated with 4 response categories. Factor loadings (λ) for the first 20 items ranged from .3 to .9 to mimic what is typically encountered in validation studies of patient-report outcome (PRO) measures. Item parameters are presented in Table 1. The 21st item was to represent a patient global impression of severity item (PGIS). Conceptually, the PGIS item should be a near perfect single-item measure of the construct. Therefore, item 21 was hard coded with a loading of .9. Values of θ were simulated at the two time points. The θ at the first time point (t_0) was $\theta_0 \sim MVN(0, \Sigma)$, with Σ a covariance matrix of 1. The second time point (t_1) applied a decrease to the distribution of theta at t_0, with $\theta_1 = \theta_0 + N(-0.6,1)$. Item response data were then simulated separately for t_0 and t_1 using the graded response model (GRM) parameterization with the common population parameters.

Table 1. Simulated item parameters for the population

Item	Factor Loading (λ)	Slope (a_1)	Intercepts			
			d_1	d_2	d_3	d_4
1	0.65	1.44	3.20	1.00	-0.93	-2.65
2	0.37	0.68	3.07	0.59	-0.87	-3.47
3	0.58	1.22	2.80	0.76	-0.96	-2.58
4	0.37	0.67	2.83	0.29	-1.28	-2.51
5	0.54	1.10	2.98	1.13	-1.46	-2.55
6	0.62	1.33	2.50	0.69	-0.43	-2.57
7	0.43	0.82	2.53	0.33	-0.47	-2.89
8	0.34	0.62	2.56	1.07	-0.87	-2.50
9	0.60	1.28	3.13	0.59	-1.08	-2.73
10	0.35	0.64	2.50	0.88	-1.47	-2.52
11	0.41	0.77	3.45	1.08	-0.96	-3.48
12	0.50	0.98	3.42	0.97	-0.74	-2.68
13	0.88	3.21	2.73	1.12	-0.52	-2.50
14	0.79	2.21	3.40	1.33	-0.06	-2.56
15	0.48	0.93	3.03	0.90	-0.26	-2.50
16	0.69	1.61	3.30	0.83	-1.01	-3.07
17	0.64	1.41	2.59	1.10	-0.93	-3.50
18	0.78	2.09	3.13	1.70	-1.04	-2.62
19	0.43	0.81	3.36	1.08	-1.19	-2.81
20	0.59	1.24	3.40	1.36	-1.23	-2.74
PGIS	0.90	3.51	2.25	0.75	-0.75	-2.25

Next, four scores were generated from the simulated response data at t_0 and t_1, respectively: 1) a sum score based on the first 20 items (SS), 2) expected a-priori (EAP) scores, 3) T-transformations of the EAP scores (TS), and 4) the true score (τ). The sum score was simply the sum of the responses to the first 20 items within t_0 and t_1. EAP scores at t_0 were generated by fitting a GRM to the responses to the first 20 items at t_0. EAP scores at t_1 were created by carrying forward the item parameters from the fitted model at t_0 to the GRM fit to the responses at t_1 and freeing the mean and variance of θ_1 to maintain factor invariance. EAP scores were then computed for t_1 responses. T-score transformations of the EAP scores were created by applying the basic transformation to the EAP scores within each time point—i.e., $\theta \sim N(0,1) \rightarrow T \sim N(50,10)$. A final true score ($\tau$) was obtained by applying the population parameters used to create the data to a GRM model fit to all 21 items simulated at each time point and freeing the mean and variance. This set of parameters was then used to generate the EAP scores for the observed responses at t_0 and t_1.

Application to PROMIS® Pain Interference

The NIH Patient Reported Outcomes Measurement Information System (PROMIS®) is an innovative set of PROs covering multiple domains in physical, mental, and social health that leverages item response

theory (IRT) [10]. Its basis in IRT allows for the construction of large item banks that represent the construct which can be tapped to implement the PRO in various ways, including computer adaptive tests (CAT). Though IRT provides a framework for understanding the performance of individual items, it also generates highly-reliable scores, especially under CAT implementation, where the most informative items from an item bank are selected in sequence until a pre-specified reliability threshold is reached [11]. These characteristics make PROMIS® a good resource to explore methods to estimate individual change.

We sourced data from the PROMIS® 1 Wave 2 Depression and Pain validation study (Protocol 07-04) to further compare the raw and regression estimators [12]. This was a prospective longitudinal study aiming to test the validity of the PROMIS Depression and Pain item banks in a "real world" setting. Among PROMIS instruments administered, the PROMIS Pain Interference adult item bank v1.1 was administered by CAT at a baseline timepoint, then again at one and three months post-baseline. Eligible patients had a diagnosis of low back pain with or without sciatica for at least 6 weeks and were scheduled for any kind of spinal injection. The PROMIS Pain Interference adult item bank was developed during PROMIS Wave 1 and contains 40 items in total focusing on the consequences of pain in the patient's life, including impacts on social, cognitive, emotional, physical, and recreational activities. All items are universal (i.e., not focused on a particular clinical population or health condition) [13]. Per PROMIS standards, an IRT score (expected a posteriori) is transformed to a T-score with a population mean of 50 and standard deviation of 10, and higher scores indicate greater pain interference.

We analyzed PROMIS pain interference T-scores for 159 patients at baseline and one month post-baseline. As the pain interference scores were based on a CAT system, no sum scores were analyzed. Raw change scores and Lord & Novick (LN) estimates were based on T-scores from the two time points and Cronbach and Furby (CF) estimates included BPI total scores from the same time points as additional information [14]. As the PROMIS® pain interference scores are based on CAT methodology and therefore do not have a set number of items, empirical reliability based on the individual standard errors was used for the reliabilities at each time point [15].

Direct comparisons between estimators were made as well as to assessments of change based on the IRT-determined standard error of individuals' T-scores—i.e., $T \sim N(50,10)$.

Statistical analyses

All analyses used R version 3.6.3 [16]. Simulated data and GRM were carried out using *mirt* [17]. Linked and unlinked correlations, mentioned in Equation 6 and Equation 7, were calculated using *CorrMixed* [18]. Reliability estimates were generated using *mirt* and *psych*, where appropriate [17, 19].

Methods for the comparison of scores

To meet the main objective of this study, several change scores (i.e., $t_0 - t_1$) were calculated using the sum scores, T-scores, and true scores (τ). Specifically for the simulated data, the raw change scores and regression estimators were calculated. Additionally, since the objective of the analyses was to compare the performance of these different methods for calculating change scores, all were placed on a common z-scale metric. Of main interest were the two regression-based as they adjust for measurement error in calculating individual change scores.

Using the simulated data, change scores and regression estimators were calculated for both SS and TS, totaling six scores for comparison to $\Delta\tau$—i.e., the true score change. Also, standardized effect sizes were used to characterize the recovery of $\Delta\tau$. To calculate these effect size differences, the absolute value of the z-transformed change scores was subtracted from the z-transformed $\Delta\tau$. These standardized differences between the calculated change score and $\Delta\tau$ can be conceptualized as effect sizes (*d*), with *d* = 0.2 considered a small, *d* = 0.5 a medium, and *d* = 0.8 a large [20]. Probability density function (PDF) curves for the standardized differences in the change scores were used to compare the recovery of $\Delta\tau$, as were descriptive statistics.

For the PROMIS® data, there is no $\Delta\tau$ and thus no comparison to a true score is possible. Also, since the data were collected in a CAT format, patients had different numbers of items answered. Thus, T-scores from the PROMIS® validation and the individual-level standard errors from the CAT scores were used for assessment of these data.

Significant change was then assessed with each of the Raw, LN, and CF estimators using the standard error of measurement (SEM) and standard error of prediction (SEP), which were used to generate a confidence interval in each data set for a lower 95% CI limit.[8, 9].

Results

Simulated Data

Details of the simulation and additional descriptive statistics for the simulation can be provided upon request. Descriptive statistics of the ΔPGIS, $\Delta\tau$, and the change score estimators appear in Table 2.

Table 2. Descriptive statistics of change scores

n = 200	Mean	Standard Deviation	Minimum	Median	Maximum
$\Delta PGIS$	-0.58	1.570	-4.00	0.00	4.00
$\Delta\tau$	-0.64	1.060	-3.29	-0.58	2.64
ΔT-score	-5.99	9.930	-31.28	-5.30	21.38
ΔT-score (LN)	-5.99	7.710	-25.44	-5.53	15.28
ΔT-score (CF)	-5.99	6.250	-22.33	-6.00	11.92
ΔSum score	-6.75	12.020	-37.00	-7.00	26.00
ΔSum score (LN)	-6.75	6.280	-21.46	-6.83	10.35
ΔSum score (CF)	-6.75	5.150	-19.39	-7.23	8.64

Although on different scales, making some direct comparisons difficult, both TS and SS showed a similar pattern for change score estimates. Across estimators, including more information in the form of reliability and/or covariates reduced variability in estimates as evidenced by smaller standard deviations for LN and CF estimates. The impact was more pronounced for the SS than the TS with the LN and CF standard deviations approximately half of the raw change scores.

To compare estimates based on a comparable scale, deviations from $\Delta\tau$ on the Z-score scale were computed, expressed as a standardized effect size, *d*, and are shown in Table 3.

Table 3. Summary of absolute deviation effect sizes (d) between estimators and $\Delta\tau$

Score Compared to $\Delta\tau$	Mean (95%CI)	|d| ≥ 0.2	|d| ≥ 0.5	|d| ≥ 0.8	Range
ΔTS	0.16 (0.142, 0.175)	31.5%	1.0%	0.0%	0.60
ΔT-score (LN)	0.17 (0.153, 0.189)	37.5%	1.0%	0.0%	0.57
ΔT-score (CF)	0.25 (0.225, 0.273)	54.5%	10.0%	0.5%	0.82
ΔSum score	0.28 (0.247, 0.314)	58.0%	15.5%	4.5%	1.49
ΔSum score (LN)	0.27 (0.242, 0.298)	57.0%	13.5%	1.0%	1.13
ΔSum score (CF)	0.35 (0.307, 0.387)	62.0%	22.0%	8.5%	1.56

Among TS estimates, similar results were observed with deviations largest for the CF estimates. The LN estimates appear to be slightly better than the raw change scores, but marginally so. For the SS, a similar pattern was observed with CF estimates showing higher levels of deviation from true scores. The small

differences seen between SS and LN for TS scores are more pronounced with 4.5% of raw changes in SS reaching large effect sizes whereas only 1% of LN estimates reaching that same level. Distributions of the deviations are displayed in Figure 1.

[Figure 1 about here]

Using the estimates to determine a generalized limit for improvement, estimates for significant improvement are presented in Table 4.

Table 4. Standard errors of measurement and prediction

Score	SEM (Lower 95% Confidence Limit)	SEP (Lower 95% Confidence Limit)
ΔTS	2.18 (-4.300)	3.05 (-6.009)
ΔT-score (LN)	1.95 (-3.840)	2.72 (-5.372)
ΔT-score (CF)	1.90 (-3.744)	2.64 (-5.210)
ΔSum score	4.06 (-7.999)	5.64 (-11.112)
ΔSum score (LN)	2.67 (-5.267)	3.73 (-7.354)
ΔSum score (CF)	4.38 (-8.645)	5.86 (-11.561)

For the TS estimates, SEM and SEP values were similar for LN and CF estimates with lower 95% CI limits indicating approximately five points indicating a significant improvement. For the SS estimates, the LN estimates showed the best precision as it showed the smallest confidence limit with an improvement of just over seven points indicating a significant improvement.

PROMIS® Pain Interference

Details of the PROMIS® data analysis can be provided upon request. Descriptive statistics for the PROMIS® scores used for the current analyses are in Table 5.

Table 5. Descriptive statistics of PROMIS® data analyzed

	Administration	n	Mean	Standard Deviation	Minimum	Median	Maximum
Pain Interference T-score	1	159	63.57	6.393	64.18	38.58	77.78
	2	158	59.49	7.557	59.50	38.58	77.78
	ΔRaw	158	-4.03	6.74	-31.93	-3.71	10.83
	ΔLN estimates	158	-4.03	3.89	-20.18	-3.59	4.39
	ΔCF estimates	158	-4.03	4.37	-21.62	-3.63	5.85
BPI Total Score	1	159	5.57	2.350	0.00	5.86	9.86
	2	158	3.77	2.710	0.00	3.36	10.00
	ΔRaw	158	-0.14	2.360	-7.86	-1.29	2.86

Reliability estimates for Administrations 1 and 2 Pain Interference T-scores were also computed using the individual T-score SE values with $r_{11'} = .924$ and $r_{22'} = .935$ for each administration, respectively. Also, linked and unlinked correlations between administrations were calculated and found to be $\rho_{12} = .828$ and $r_{12} = .834$. These calculations then fed into the assessment of reliability of the change in Pain Interference T-score with $r_{DD'} = .594$.

The reliabilities and correlations were then employed in calculating the LN and CF estimates, also summarized in Table 5. The LN and CF estimates showed notably smaller standard deviation, minima and maxima as expected with regression estimators.

Comparison of the raw and regression estimates, as in the simulated analysis, were done on a Z-score and Figure 2 shows the density plots of those estimates. All three estimates are very close to one another in distribution, indicating that the one with the smallest dispersion on its natural scale is likely the strongest estimator.

[Figure 2 about here]

Table 6 displays the computed standard errors. To ground these estimates, descriptive statistics of the individual-level T-score standard errors was also computed as this was the natural error measurement for the CAT scores in the dataset.

Table 6. Standard error estimates for PROMIS® Pain Interference difference estimates

Estimator	N	SEM	Lower 95% Confidence Limit	SEP	Lower 95% Confidence Limit
ΔRaw	158	4.30	-8.49	5.42	-10.71
ΔLN estimates	158	2.48	-4.90	3.13	-6.19
ΔCF estimates	158	2.78	-5.50	3.51	-6.94

		Mean	Lower 95% Confidence Limit	Minimum	Median	Maximum
Individual T-score SE	158	1.93	-3.82	1.55	1.78	5.61

The lower confidence limit values for each estimator indicate that the LN estimates have the most precision and are closest to the Individual SE summaries, which take the most information into account.

Discussion

The current paper's objective was to illustrate how the use of additional information can enhance the assessment of raw change scores, especially for individuals. We presented methodologies that, although seasoned, are not commonly used in PRO measure applications. The LN and CF estimators both incorporate measurement error by using reliabilities of the measurements at two occasions as well as the reliability of the difference between those occasions into account. The CF estimator also adds additional information in the form of one or more covariates at one or both time points.

With regard to the simulated dataset, within score type the pattern was clear that the regression estimators reduced dispersion with smaller SD values and minima and maxima closer to the mean in both LN and CF cases. Attenuation of dispersion for T-scores was apparent, but noticeably more so for the Sum scores. Whereas one could argue that for the T-score estimates, there were small differences, the Sum score comparison showed that SD values for the regression estimates reduced the Raw difference SD by close to half. Comparisons of deviances from τ values on the Z-scale and as *d*-values also showed a similar pattern, although the CF estimator showed a relatively high percentage of individual deviances at a medium or high *d*-value for both T- and Sum scores. This result warrants further investigation, but may be related to use of a single item PGIS as the grading of this variable compared to the continuous scores could affect the estimator.

Specifically with regard to the Sum scores, the Z-scores showed poorer, i.e., larger, deviances from the τ values in general with both Raw and CF showing 4.5% and 8.5% of *d*-values at 0.8 or higher. Comparably, the LN estimator was only 1%. While further investigation is warranted, these results suggest that, especially for sum scores, that the LN estimator performs best.

Comparisons of PROMIS® Pain Interference scores also showed attenuated dispersion for the regression estimators with the LN estimator showing the smallest values, indicating better precision. The right panel of Figure 2 underscores this point as plots of the estimators on the Z-scale reveals very small distributional differences, therefore, all things equal, the LN would be the estimator of choice with the smallest dispersion.

Using the individual SE values from the CAT T-scores as a basis for further comparing the performance of the estimators suggested the regression estimators had the closest values of SEM and SEP as compared to the mean SE of the CAT-determined errors. Further, when the lower 95% confidence limits for the regression estimators are computed, they are remarkably close to the maximum value of the Individual

SE. This would suggest that both SEP-based limits for individual improvement, in the current case, would catch classify improvement for 95% of those that would exceed a value of improvement based on their individual SE value, i.e., $1.98 \times SE$.

While it is best, if available, to use the tools available from IRT analyses to determine individual changes in latent states over simply looking at raw score values, we feel that using the regression estimators described here present a compromise that is available in many situations including those for which a legacy instrument has been validated with a CTT, e.g., sum score, set of techniques.

Conclusions

Our results suggest that calculating change using regression estimates may result in more increased measurement sensitivity. Both regression estimators incorporate information other than baseline scores, such as measurement error and the correlation between scores at different time points, into the estimation of a change. Using these scores in lieu of raw differences can help better identify individuals who experience real underlying change in PROs in the course of a trial, and enhance the established methods for identifying thresholds for meaningful within-patient change in PROs. Further, the use of regression estimators for change may result in increased power to detect change in trials.

Of note, the CTT estimators we have explored here still contain an element of marginalization with regard to the information contained in PRO item responses. The EAP or other scoring methods employed for scoring IRT models allow for individual-level errors to be calculated, based on the respondents' levels on the latent variable being measured—i.e., θ. While we advocate using IRT when appropriate, we also think the regression estimators presented here represent a better and more accessible alternative to raw change scores for determining individual improvement or worsening.

List of Abbreviations

CAT	Computerized Adaptive Testing
CF	Cronbach & Furby (complete estimator)
CTT	Classical Test Theory
EAP	*Expected a-priori*
GRM	Graded response model

IRT	Item-Response Theory
LN	Lord & Novick
MVN	Multivariate normal distribution
PRO	Patient-reported outcome
PROMIS®.	Patient-Reported Outcome Measurement Information System
SE	Standard Error
SEM	Standard Error of Measurement
SEP	Standard Error of Prediction
SS	Sum Score
TS	T-Score

References

1. U.S. Food and Drug Administration (2019) Patient-focused Drug Development Guidance Public Workshop - Discussion document: Incorporating clinical outcome assessments into endpoints for regulatory decision-making. 1–50
2. Coon CD, Cook KF (2018) Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 27:33–40
3. Kim-Kang G, Weiss DJ (2008) Adaptive measurement of individual change. Zeitschrift für Psychol Psychol 216:49–58
4. Lord FM (1958) Further problems in the measurement of growth. Educ Psychol Meas 18:437–451
5. Lord FM (1956) The measurement of growth. ETS Res Bull Ser 1956:i--22
6. McNemar Q (1958) On growth measurement. Educ Psychol Meas 18:47–55
7. Cronbach LJ, Furby L (1970) How we should measure change--or should we? Psychol Bull 74:68–80
8. Lord FM, Novick MR (1968) Statistical Theories of Mental Test Scores. Addison-Wesley Pub. Co, Reading, MA

9. Cascio WF, Kurtines WM (1977) A practical method for identifying significant change scores. Educ Psychol Meas 37:889–895 . https://doi.org/10.1177/001316447703700411

10. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, Dewalt D, Fries JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R (2010) The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 63:1179–1194 . https://doi.org/10.1016/j.jclinepi.2010.04.011

11. Segawa E, Schalet B, Cella D (2020) A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. Qual Life Res 29:213–221

12. Pilkonis P (2016) PROMIS 1 Wave 2 Depression (V1 ed.)

13. Amtmann D, Cook KF, Jensen MP, Chen W-H, Choi S, Revicki D, Cella D, Rothrock N, Keefe F, Callahan L, Lai J-S (2010) Development of a PROMIS item bank to measure pain interference. Pain 150:173–182

14. Cleeland CS, Ryan KM (1994) Pain assessment: global use of the Brief Pain Inventory. Ann. Acad. Med. Singapore 23:129–138

15. Samejima F (1994) Estimation of reliability coefficients using the test information function and its modifications. Appl Psychol Meas 18:229–244

16. R Core Team (2020) A Language and Environment for Statistical Computing. R Found. Stat. Comput. https://www.R--project.org

17. Chalmers RP (2012) mirt: A Multidimensional Item Response Theory Package for the R Environment. J Stat Softw 48:1–29 . https://doi.org/10.18637/jss.v048.i06

18. der Elst W, Molenberghs G, Hilgers RD, Verbeke G, Heussen N (2016) CorrMixed: Estimate Correlations Between Repeatedly Measured Endpoints (Eg, Reliability) Based on Linear Mixed-Effects Models. R package version 0.1--13

19. Revelle W (2020) psych: Procedures for Psychological, Psychometric, and Personality Research

20. Cohen J. (1988) Statistical Power Analysis for the Behavioural Science (2nd Edition). In: Statistical Power Anaylsis for the Behavioral Sciences (2nd Edition). Taylor & Francis Group

Declarations

Ethics approval and consent to participate

Not applicable

Consent for Publication

Not applicable

Competing interests

The authors have no competing interests to declare

Funding

Not applicable

Authors' contributions

All authors contributed to the conceptualization, drafting, and review of the manuscript. DAA conducted the analyses. JDP supplied the PROMIS® dataset. All authors approved the final manuscript.

Availability of data and materials

R scripts to generate the simulated data are available on request.

The PROMIS® 1 Wave 2 Pain Depression dataset can be requested here:
https://doi.org/10.7910/DVN/ZDIITC

Acknowledgements

Not applicable

Figures

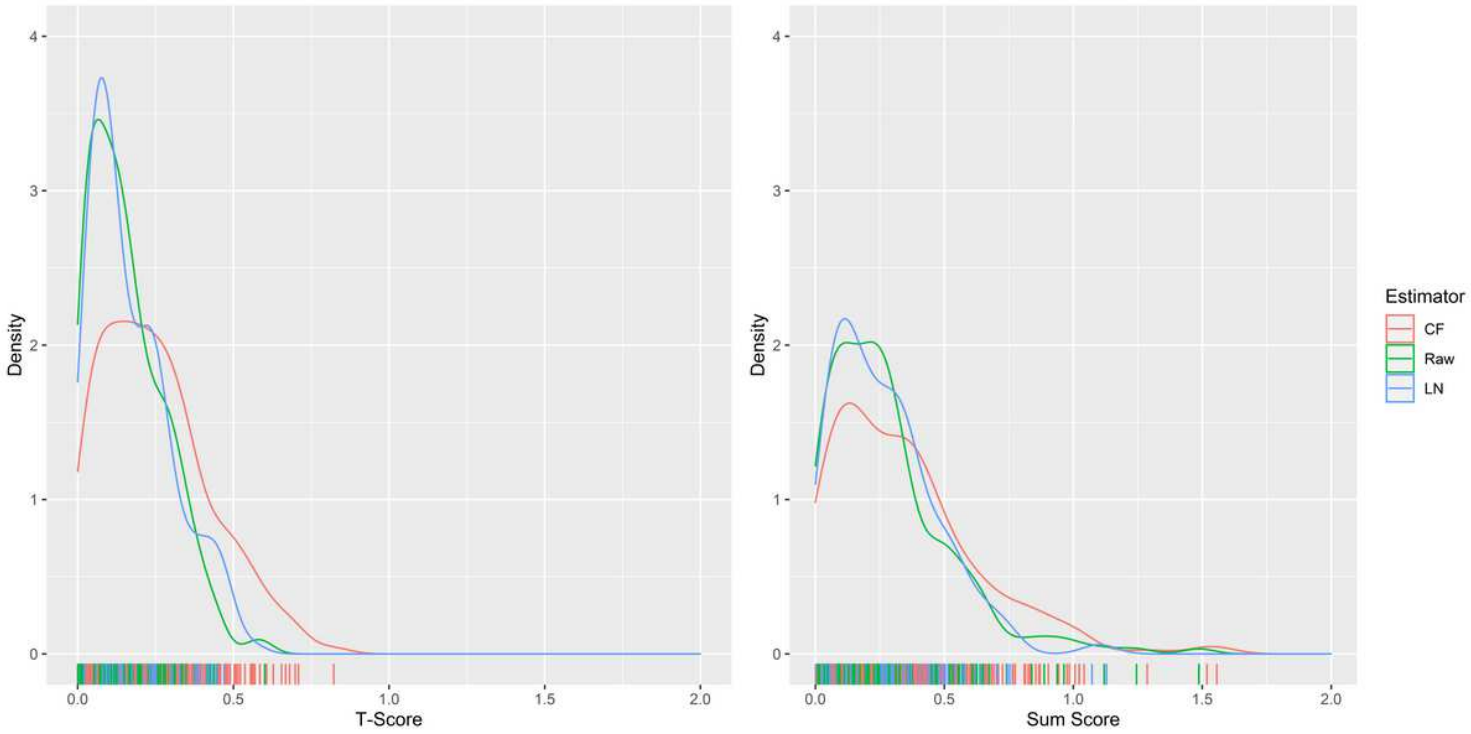

Figure 1

Distributions of the absolute deviations of the Raw, LN, and CF estimators from the simulated true score (τ). The left panel displays T-scores and the right displays Sum scores.

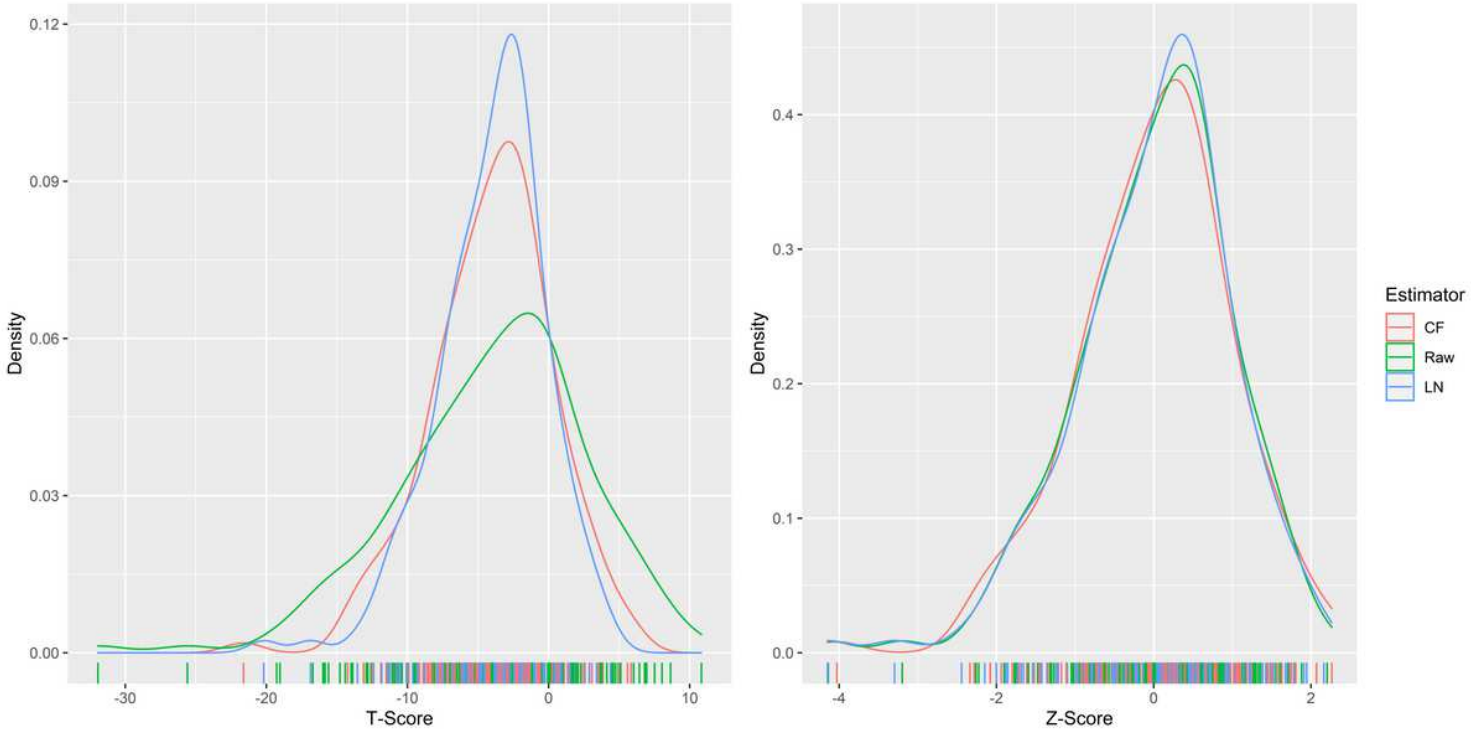

Figure 2

Probability density plots of PROMIS® Pain Interference change score estimates. The left panel represents the T-score and the right panel plots the equivalent Z-scores.