Published in final edited form as:

Qual Life Res. 2018 June ; 27(6): 1589–1597. doi:10.1007/s11136-018-1817-4.

Longitudinal and Dynamic Measurement Invariance of the FACIT-Fatigue Scale: an application of the Measurement Model of Derivatives to ECOG-ACRIN study E2805

Ryne Estabrook,

Department of Medical Social Sciences, Northwestern University, Chicago, IL

David Cella,

Department of Medical Social Sciences, Northwestern University, Chicago, IL

Fengmin Zhao,

Dana Farber Cancer Institute - ECOG-ACRIN Biostatistics Center, Boston, MA

Judith Manola,

Dana Farber Cancer Institute - ECOG-ACRIN Biostatistics Center, Boston, MA

Robert S. DiPaola,

University of Kentucky, Lexington, KY

Lynne I. Wagner, and

Wake Forest University Health Sciences, Winston-Salem, NC

Naomi Haas

University of Pennsylvania, Philadelphia, PA

Abstract

Purpose—While quality of life measures may be used to assess meaningful change and group differences, their scaling and validation often rely on a single occasion of measurement. Using the 13-item FACIT-Fatigue questionnaire at three timepoints, this study tests whether individual items change together in ways consistent with a general fatigue factor.

Methods—The Measurement Model of Derivatives (MMOD) is a novel method for measurement evaluation that directly assesses whether a given factor structure accurately describes how individual test items change over time. MMOD transforms item-level longitudinal data into a set of orthogonal change scores, each one representing either a within-person longitudinal mean or a different type of longitudinal change. These change scores are then factor analyzed and tested for invariance. This approach is applied to the FACIT-Fatigue scale in a sample of patients with renal cell carcinoma treated on 'ECOG-ACRIN Cancer Research Group (ECOG-ACRIN) study 2805.

Results—Analyses revealed strong evidence of unidimensionality, and apparent factorial invariance using traditional techniques. MMOD revealed a small but statistically significant difference in factor structure ($\Delta\chi^2_{12} = 49.597, p < .001$), where factor loadings were weaker and more variable for measuring longitudinal change.

Conclusions—The differences in factor structure were not large enough to substantially affect scale usage in this application, but they do reveal some variability across items in the FACIT-Fatigue in their ability to detect change. Future applications should consider differential sensitivity of individual items in multi-item scales, and perhaps even capitalize upon these difference by selecting items that are more sensitive to change.

Keywords

Measurement invariance; longitudinal modeling; factor analysis; fatigue

Introduction

Fatigue is a common and important symptom associated with a variety of medical conditions, including cancer. Fatigue is familiar to healthy individuals, and varies in the general population as well as across various diseases (Cella, Lai, & Stone, 2011). Fatigue can be caused by several physical and mental disorders, and it can be a side effect of treatment. As a result, it is often included in longitudinal studies of treatment outcomes.

A variety of fatigue measures exist, allowing researchers to select measures that are either specific to a particular disease or more general measures. The FACIT-Fatigue scale is a thirteen-item self report scale that has undergone extensive psychometric validation (Yellen, Cella, Webster, Blendowski, & Kaplan, 1997). FACIT-Fatigue scale scores have been referenced to cancer patients and to the general US population, and it has been validated in a variety of disease states, including rheumatoid arthritis (Cella et al., 2005), anemia (Acaster et al., 2015), Parkinson's disease (Hagell et al., 2006), systemic lupus erythematosus (Kosinski, Gajria, Fernandes, & Cella, 2013), as well as multi-morbidity assessment spanning cancer, stroke, and HIV (Butt et al., 2013). It has also been validated in multiple languages (Kwakkenbos et al., 2014; Dapuerto, Del Carmen Abreu, Francolino, & Levin, 2014).

Symptoms like fatigue demonstrate both inter-individual (between-person) and intra-individual (within-person) variation, including variation over time. This variation over time is what is of most important to capture in a clinical trial, as this variation corresponds to the effect of treatment. Thus, it is important that measures of fatigue not only reliably measure individual differences in overall level at one point in time, but also measure short- and long-term patterns of change.

Unfortunately, most psychometric testing relies on only a single observation per person, defining constructs in terms of what distinguishes individuals from one other. Even when longitudinal data are used for measurement evaluation, many existing methods treat additional measurement occasions separately using multiple group approaches (Widaman, Ferrer, & Conger, 2010; Jabrayilov, Emons, de Jong, & Sijtsma, 2017), ignoring any information about individuals change. This level of testing may be sufficient for trait-like constructs where interindividual differences are of interest. Oort (2005) has demonstrated the utility of SEM to separate construct-level change from changes in item characteristics, highlighting the utility of longitudinal data for this purpose. Similarly, Pentz & Chou (1994)

have demonstrated that group-based invariance studies comparing treatment and control groups can be accomplished with two-occasion longitudinal data.

However, longitudinal data are different than group-based approaches as they reveal information about change that can be attributed to specific individuals rather than samples or populations. Molenaar (2004) has pointed out that the structures of between-person and within-person changes often differ, and that dynamical and within-person analyses are required to move from population description to the study of mechanisms and processes. Estabrook (2015) showed via simulation how measures can appear unidimensional and invariant across timepoints despite more complex causes for within-person item-level change. Dynamical longitudinal models can inform stricter invariance tests than existing approaches and can help build measures that are sensitive to small within-person change.

The goal of this paper is to evaluate the structure of the FACIT-Fatigue scale, specifically whether the items that make up the FACIT-Fatigue show a coincident change consistent with changes in a general fatigue factor.

Methods

Sample

Data for this analysis come from a randomized clinical trial conducted by the 'ECOG-ACRIN Cancer Research Group (ECOG-ACRIN), comparing two different VEGFR tyrosine kinase inhibitors (sorafenib and sunitinib) each to placebo, in people with early stage renal cell carcinoma (E2805; ClinicalTrials.gov identifier: NCT00326898). Patients were randomly assigned to either placebo or one of two treatment arms, with dosage of those treatments varying over the course of the study as part of the experimental protocol (Haas et al., 2016). All patients were measured at baseline, 10 weeks, and 22 weeks post treatment.

The quality of life (QoL) component of the study was added after enrollment began. Based on a power analysis, of the 1943 patients enrolled in this trial, 418 participated in the QoL substudy and gave data at baseline: 45 patients were invited to the QoL arm but did not complete baseline data collection. All 418 individuals and all three timepoints of QoL data were included in these analyses. 37 patients did not participate at the 10 week followup (n=381), and 92 patients did not participate at the 22 week followup (n=326).

More information about this trial, including details about sampling, research design, and results for each study arm, can be found in Haas, et al (2015).

Measure

The FACIT-Fatigue is a 13-item scale that assesses fatigue in everyday life (Yellen et al., 1997). Individuals respond on a 5-point Likert scale, indicating the degree to which they have experienced the listed symptoms over the past week. The five response categories are "not at all", "a little bit", "somewhat", "quite a bit", and "very much". Higher scores indicate less fatigue. The individual items in the FACIT-Fatigue scale are listed in Table 3.

Modeling Approach: The Measurement Model of Derivatives

This paper will utilize a novel approach to longitudinal invariance, as our interest is not in whether the FACIT-Fatigue scale has the same structure at each timepoint, but whether the scale is a unidimensional and invariant measure of *change*. This can be accomplished using the Measurement Model of Derivatives (MMOD; Estabrook, 2015), a factor analytic approach for longitudinal measurement invariance. This approach uses generalized orthogonal local derivatives (GOLD; Deboeck, 2010) to differentiate item-level data prior to a measurement model. Put more simply, MMOD transforms item-level longitudinal data into item-level change scores, then fits factor models to make sure that the factor structure of these change scores matches the factor structure of between person differences.

MMOD's item-level differentiation can be done either with a latent variable model or through a data transformation. We will focus on the transformation method in this paper, which is akin to calculating change scores with more than two timepoints. This can be best explained through example: the zeroth, first, and second derivatives of some variable y (denoted y , $\dot{y}$, and $\ddot{y}$) can be calculated as linear combinations of three timepoints of data (y_1 , y_2 , and y_3). The zeroth derivative (y) represents position, and can be calculated as simply the mean across the three occasions of measurement. The first derivative represents velocity or rate of change, and can be calculated as the (linear) change over the observed interval: with three occasions, this is simply the change from the first timepoint to the last divided by the elapsed time ($\frac{y_3 - y_1}{2}$). Finally, the second derivative represents acceleration, or how much the first derivative is changing. This can be calculated by calculating two rates of change ($\dot{y}_1 = y_2 - y_1$, $\dot{y}_2 = y_3 - y_2$), then calculating the difference between those two rates of change ($\ddot{y} = \frac{\dot{y}_2 - \dot{y}_1}{2}$).

While three dimensional transformations can be done by hand, MMOD and GOLD derivatives are more commonly done by compiling these equations into a matrix of weights, then multiplying data by that weight matrix to yield derivatives. Equation 1 lays out the algebra for this three-dimensional transformation. A weight matrix W contains three columns, each representing the weights for combining y_1 , y_2 , and y_3 to create three orders of derivatives. In Equation 2, a vector or row of longitudinal data is transformed into a vector or row of derivatives through multiplication by W . Alternatively, entire covariance matrices can be transformed using this method. Equation 3 shows how a covariance matrix may be pre- and post-multiplied by a weight matrix to yield a covariance matrix of derivative transformed data.

$$W_{3 \times 3} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & -1 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad (1)$$

$$[y_1 \ y_2 \ y_3]W_{3 \times 3} = [y \ \dot{y} \ \ddot{y}] \quad (2)$$

$$W'_{3 \times 3}Cov(y)W_{3 \times 3} = Cov(D) \quad (3)$$

These derivative-transformed data can then be factor analyzed to assess the structure of change, with sets of factors predicting each order of item derivative in the exact same way factors at timepoints are fit to longitudinal data. Factors fit to zeroth derivatives will indicate the structure of between-person differences in overall level or mean: in this analysis, zeroth derivative factors would indicate whether people who show high levels on one item also show high levels on all other fatigue items. First derivative factors would model the associations between changes on all items over the entire study interval, testing whether individuals who increase on one item show coincident increases on other items. Second derivative factors would test associations between acceleration and deceleration, testing whether people who change a lot early then slow down (or change little early in the study and more later) do so across all items. These two change factors can be interpreted together or individually, and tested for invariance like any other factor model.

This method bears great resemblance to McArdle's (McArdle, 1988) factor-of-curves (FOCUS) model, which fits factor analyses to item-level growth curves. MMOD differs from FOCUS in that it relies not on growth curve models, which require researchers to select and fit growth curves for each item, but on the calculation of orthogonal change scores. The change scores central to MMOD are full rank, meaning that there are as many change scores as timepoints, so no misfit can be created. These scores are also orthogonal in the same way that planned contrasts in ANOVA or regression are orthogonal, further preventing longitudinal misfit. MMOD allows for the analysis of the measurement structure underlying item-level change without requiring an explicit hypothesis or longitudinal model for that change, instead focusing on measurement evaluation and leaving hypotheses about change on the ensuing construct for future analyses. While derivative transforming data is slightly more complex than analyzing each timepoint individually, the GOLD method underlying MMOD allows for analyzing each order of derivative similar to a cross-sectional dataset while providing useful information about item-level change.

Data Processing

Analyses were completed using summary data. A polychoric correlation matrix was calculated from observed longitudinal data, providing maximum likelihood estimates of correlations between ordinal variables. This correlation matrix was then derivative transformed using the weight matrix and algebra in Equations 1–3, yielding a covariance matrix of item level zeroth, first, and second derivatives. This covariance matrix can be analyzed as a unit in the same way the longitudinal polychoric correlation matrix was. Cross-sectional models at each timepoint and dimensionality tests for single orders of derivatives were completed by subsetting these raw and derivative transformed polychoric

correlations and selecting only relevant variables (e.g., all items at T1 or all items corresponding to zeroth derivatives).

Subsets of our models were refit using full information maximum likelihood for ordinal data, and again treating all outcomes as continuous, as a method of testing the importance of modeling assumptions and missingness. This modeling approach provides asymptotically unbiased parameter estimates in the presence of missing data and attrition, but places significant restrictions on model estimation. In OpenMx (Neale et al., 2015), ML requires twenty or fewer dimensions of integration, effectively capping all analyses at twenty categorical variables. As such, we replicated single occasion models (with thirteen items and thus thirteen categorical variables) for ordinal data and complete longitudinal models with continuous data, finding minor differences in parameters that do not materially affect parameter interpretation or change significance tests in any way. Based on the robustness of our findings with varying assumptions, we are confident in the accuracy of our polychoric results despite the associated problems with the use of a single polychoric correlation matrix.

Analytic Plan

Measurement evaluation of the FACIT-Fatigue took place in several parts. We first assessed cross-sectional fit and dimensionality using both traditional (i.e., single timepoint) and derivative-transformed methods to assess whether levels and changes of the individual items are consistent with a single latent construct. We then tested for factorial invariance over time using Meredith's (Meredith, 1964a, 1964b, 1993) levels of factorial invariance. These tests were first fit to all three timepoints longitudinally, then to all orders of derivatives. Finally, loading patterns were compared across all orders of derivative transformed data to assess differences between level and change. Latent variables were constrained to unit variance for identification. All common latent variables had freely estimated correlations in longitudinal models. Models with invariance constraints made the identification constraint at the first timepoint, then freely estimated factor variances as subsequent timepoints. All longitudinal models included freely estimated autocorrelations for each item over time to prevent correlated errors and other item-specific longitudinal trends from impacting results. Models will be compared via likelihood ratio test, with conventional SEM fit statistics (CFI, TLI, RMSEA) used to evaluate overall fit.

Results

Cross-Sectional Analyses and Dimensionality Testing

Cross-sectional models and dimensionality testing exclusively used the polychoric correlation and derivative transformed covariance matrices, both of which were separated into individual timepoints or derivatives for this portion of the analysis. Our first step was to fit a single factor model to all items at the first timepoint. This model showed poor fit at all three timepoints, as shown in the top half of Table 1. As two of the thirteen items were positively worded and may show associations owing to different response styles, we further tested the addition of a residual correlation to account for this method effect. Adding a residual correlation between the two positively worded items (AN5 & AN7) significantly

improves fit (all $\chi^2_1 > 80$, all $ps < .001$). CFI and TLI improve to approximately 0.90, which exceed some recommendations but not others (Kline, 2005). RMSEA is still poor, but can show poor fit values with few variables or subjects (Kenny, Kaniskan, & McCoach, 2015).

Given the discordance between these poor fit statistics and previous validation of this scale, we then proceeded with dimensionality testing to make sure that the theorized single factor structure is appropriate. Dimensionality was assessed by eigenvalue decomposition or scree test, the results of which are shown in Figure. Each timepoint and order of derivative showed strong evidence of unidimensionality. Results for each timepoint are shown in the left side of Figure, while results for derivative transformed data are shown on the right. All timepoints and derivatives show a strong general factor at least five times larger than the second eigenvalue (range = [5.25, 11.80]), far surpassing Ahn & Horenstein's (2013) recommendations on eigenvalue ratios. Analyses of individual timepoints show no other eigenvalues greater than 1.0, surpassing both the Kaiser rule (Kaiser, 1960) and parallel analysis criteria (Horn, 1965). Derivative-transformed data show slightly higher second eigenvalues for the first and second derivatives (first derivative: $\lambda_2=1.18$, second derivative: $\lambda_2=1.20$), suggesting that zeroth derivatives (and between person differences in level) show stronger evidence of unidimensionality than within-person change. However, these first and second derivative eigenvalues each fall short of both median ($\lambda=1.24$) and 95th percentile ($\lambda=1.31$) parallel analysis criteria for this sample size, and are closer to the 5th percentile ($\lambda=1.18$).

Put together, these results show no evidence for a second fatigue factor, but poor fit of cross-sectional models indicate that a single factor may not be appropriate.

Invariance Testing Over Time

Next, we subjected traditional longitudinal data to invariance testing over time and order of derivative. As the polychoric correlation and derivative-transformed covariance matrices used in these analyses don't have item intercepts required for strong and strict factorial invariance, we present tests of weak factorial invariance that constrain factor loading patterns over time or derivative order. Results for both traditional longitudinal and derivative transformed tests of invariance are presented in Table 2. All longitudinal models further include autocorrelations for each item as recommended by Oort (2005) to reduce misspecification of the longitudinal trend, as this misspecification would generate additional misfit and possibly impair evaluation of measurement structures.

Longitudinal assessments in Table 2 show no loss of fit when factor loading patterns were constrained to invariance ($\Delta\chi^2_{24} = 29.608$, $p=.198$) and slight improvements in all fit indices. However, applications of MMOD show significant differences between constrained and unconstrained models ($\Delta\chi^2_{24} = 64.999$, $p<.001$). An intermediary model where the first and second derivatives were constrained to invariance (but differ from the zeroth derivative) shows the best fit overall, significantly improving on the fully invariant model ($\Delta\chi^2_{12} = 49.597$, $p<.001$) while showing no loss of fit relative to the fully free model

($\Delta\chi^2_{12} = 15.402, p=.220$). Together, these results show significant differences between the structure of between-person differences and the structure of within-person change.

Comparing Factor Structures for Level and Change

The previous section found a small but significant improvement in fit when zeroth derivative factors, representing between-person levels or means, had different factor loadings than the factors representing different types of change. Table 3 shows the standardized loadings for the zeroth, first, and second derivative factors, as well as T1 of the longitudinal analysis; a visual representation of the zeroth and first derivative loadings is also presented in Figure 2. These loadings are standardized from the raw output of the “MMOD Zeroth vs. Change” model from Table 2, which constrains the raw values of the first and second derivatives to equality. The standardized loadings are largest for the zeroth derivative factor and smallest for the second derivative factor, reflecting the fact that there is more variance between people than within people.

In addition to the overall shift in magnitude, several items that show moderate loadings for zeroth derivatives show weak associations with higher derivatives. In particular, item AN7 “I am able to do my usual activities” drops from a standardized loading of .696 for zeroth derivatives to .405 and .359 for the first and second derivatives, respectively. Overall, items show weaker loadings on higher order factors, and the decrease is larger in both percentage and raw magnitude for items with weaker loadings on the T1 and zeroth factors.

Discussion

The analyses presented in this paper assessed the invariance of the FACIT-Fatigue scale within an existing randomized controlled trial in oncology. Both traditional and novel methods (i.e., MMOD) were used to verify the structure of this scale in a population with renal cell carcinoma. Results showed that the FACIT-Fatigue scale was clearly unidimensional at all timepoints and over all types of change. Tests of factorial invariance showed no differences using traditional methods, but significant differences in scale structure for the assessment of change. These differences were very small and predominantly affected the weakest loading items on the scale. While these results will not strongly affect scale usage or scoring in this sample, they point to differential change in items that may affect scale construction and usage in other samples.

Single timepoint analyses and dimensionality tests show few issues with the scale. Fit is poorer than expected, but is improved by the inclusion of a residual correlation between the two positively worded items. The use of derivative-transformed data does not affect the number of factors underlying this scale. While it is possible for there to be varying numbers of factors across orders of derivatives (Estabrook, 2015), no evidence of this was found in this sample.

Invariance tests showed varying results depending on the method used. Traditional longitudinal invariance models showed no differences by timepoint, indicating that the factor structure at T1 was not significantly different than the structure at T2 or T3. However, this approach is not an explicit test of the structure of change: if persistent differences exist

between people, then people who score highly at T1 will also score highly at subsequent timepoints. To the extent that these individual differences are stable, then the longitudinal version of this model is simply replicating between-person differences at different occasions. Additionally, this model ignores item-level longitudinal change that is not caused by the factors: adding a simple autocorrelation parameter for each item significantly improves fit, indicating that there is item-level variation that is not well captured by a common trait. However, alternative longitudinal structures could be fit for both the common and unique factors, and simultaneously assessing the fit of a measurement structure and item-level longitudinal change creates difficult problems that MMOD was designed to address.

Derivative-transformed methods show a more complicated story. There are significant differences in factor loading structures, specifically differences between the zeroth derivatives and the set of higher derivatives. However, these differences are relatively small: they have little impact on model fit as measured by CFI, TLI, and RMSEA, and primarily affect items with weak loadings that have smaller impact on the fatigue trait regardless of whether between-person or within-person structures are of interest. As a further check, we generated trait scores at T1 using the standardized loadings from both the zeroth and first derivative values shown in Table 3: the scores generated from these two similar factor structures correlate greater than .99. While these differences are large enough for statistical significance and show item-level differences in sensitivity to change, these differences are not large enough to have noticeable effects on this application of the FACIT-Fatigue scale.

MMOD provided a look at these data not available with other techniques. The first and second derivatives of each item directly assessed the changes those items showed over the course of this study. More traditional methods simply assess the between-person structure at various occasions of measurement. Our application of these traditional methods revealed a common problem: each item had some type of autocorrelative structure independent of the fatigue trait, which negatively impacted fit. In our application, the scale was so strongly unidimensional that there was no reason to consider additional factors to improve fit. However, were the dimensionality tests to be less clear, longitudinal models that ignored this autocorrelative structure may have sufficiently poor fit to justify the inclusion of unnecessary common factors to account for this misspecification. MMOD was not affected to nearly the same degree by this problem: model fit was improved by the addition of a single correlation between item derivatives, but that improvement in fit had little to no effect on CFI, TLI, or RMSEA, or the factor loading patterns. As the item level derivatives span multiple timepoints, they partially model these item-level dynamics and incorporate them into latent trait assessment.

Other aspects of MMOD were not discussed for this application, but could be useful in replications and extensions of this analysis. Only three occasions of measurement are available in this study, but the derivative transformation method underlying MMOD can be extended to any number of timepoints. More complicated residual structures could be included, as could uneven spacing between observations (observations in this study were 10 and 12 weeks apart). Alternative derivative transformation methods, including those using functional data analysis (Ramsay, Hooker, & Graves, 2009), may be more useful for physiological applications with known longitudinal patterns. Additionally, Estabrook (2015)

noted that zeroth derivatives can contain some within-person information, as a within-person mean could differ from their “true” mean if measurement error or within-person variability leads to higher or lower estimates of the mean. In this example, a person could have elevated fatigue on the days surrounding measurement by chance, which would lead to inflated estimates of between-person mean and the mixture of between-person and within-person information.

This study does have limitations. As an RCT, change in this sample is likely to be induced not by a developmental process, but by the duration of treatment and possibly disease recurrence. As such, changes in individual items may not generalize to different treatments or populations. Fatigue in this renal cell carcinoma sample is worse than in the general population, with a mean score at baseline of 42.5 (SD=9.0) using the PROMIS metric lower than the general population’s mean score of 50. This sample may further have different between-person variation than other disease-specific or normative samples. Also, the use of three timepoints of measurement with relevant questionnaire data¹ over a five month interval restrict the shape and span of change that can be observed. Specifically, this design cannot detect any changes that occur over longer time periods than five months or any changes that happen over shorter periods than our measurement interval. The small differences observed over this interval may not apply to either very short-term fluctuations in fatigue (i.e., daily or weekly), or to longer maturational or developmental changes spanning years. Additionally, fit indices for the model as a whole are suboptimal. There is no evidence that the scale is multidimensional, but even the single occasion models (see Table 1) show poor fit by RMSEA in particular. Absent the addition of more factors or residual correlation parameters, fit could be improved by changing the scale itself through item deletion or addition. However, the FACIT-Fatigue scale has been extensively validated in other samples, and this poor fit could have been related to unexpected item correlations specific to this sample or population. Finally, MMOD is a novel technique and only one way of assessing change. Alternative models for the structure of item level change are possible, and should be developed and considered in the future. Further extensions of MMOD, including faster optimization to more easily handle many items and timepoints of raw data, would also aide use.

The differences between level and change in this sample were relatively small, but differences in other scales and samples could be larger and require changes to scoring methods and test construction. A scale could be good at picking up between-person differences (i.e., its good at measuring whether a person is generally fatigued), but not sensitive to within-person change (i.e., the change in fatigue due to treatment doesnt lead to improved scores). Alternatively, individual items may not change, and thus have zero loadings on change factors despite being good indicators of trait fatigue. Finally, people could show unidimensional differences in overall level (people are more or less fatigued), but a treatment or development could lead to multidimensional changes where some people increase on one subset of items, while others increase on a different subset. This would manifest as a unidimensional zeroth derivative factor, and multidimensional higher

¹Physiological and biological measures were more frequently assessed.

derivative factors. All of these cases can and should be tested for, and the knowledge about item-specific differences in sensitivity to change can help researchers build more sensitive and dynamic measures that provide maximum power to test treatment effects and model development.

The analyses presented here are a first step towards more nuanced testing of the sensitivity of measures like the FACIT-Fatigue to within-person change. While the differences found in this analysis were not large enough to affect how the scale should be used, they were statistically significant and reflect item-level differences in sensitivity to change beyond what would be expected or assumed based on more conventional modeling. To the extent that researchers are interested in change, whether due to a treatment, experimental manipulation, or developmental change, analysis like MMOD should be carried out to make sure that measures are made up of items that change together. Traditional measurement evaluation assesses whether people can be distinguished from one another. In most applications, between-person differences are much larger than within-person differences, and thus are easier to detect and assess with simple factor structures. Methods that use test-retest reliability to select reliable items will select items that are less sensitive to short-term change, further increasing sensitivity to between-person differences at the expense of within-person variation. By actively assessing and looking for items that are sensitive to change, we can build better measures that respond to treatment, improve power, and better reflect the processes we are trying to study.

Acknowledgments

This study was coordinated by the ECOG-ACRIN Cancer Research Group (Robert L. Comis, MD and Mitchell D. Schnall, MD, PhD, Group Co-Chairs) and supported by the National Cancer Institute of the National Institutes of Health under the following award numbers: CA180820, CA180794, CA189828. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government. Drs. Estabrook & Cella are supported by U02C-CA186878-01 (2014–2018) from the National Cancer Institute, National Institutes of Health.

Conflict of Interest: No authors have declared any conflicts of interest that would affect this work.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

- Acaster S, Dickerhoof R, DeBusk K, Bernard K, Strauss W, Allen LF. Qualitative and quantitative validation of the FACIT-fatigue scale in iron deficiency anemia. Health and quality of life outcomes. 2015; 13:60. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4434873&tool=pmcentrez&rendertype=abstract>. doi: 10.1186/s12955-015-0257-x [PubMed: 25980742]
- Ahn SC, Horenstein AR. Eigenvalue Ratio Test for the Number of Factors. *Econometrica*. 2013; 81(3): 1203–1227. Retrieved from <http://onlinelibrary.wiley.com/doi/10.3982/ECTA8968/abstract>. DOI: 10.3982/ECTA8968
- Butt Z, shei Lai J, Rao D, Heinemann AW, Bill A, Cella D. Measurement of fatigue in cancer, stroke, and HIV using the Functional Assessment of Chronic Illness Therapy - Fatigue (FACIT-F) scale.

- Journal of Psychosomatic Research. 2013; 74(1):64–68. DOI: 10.1016/j.jpsychores.2012.10.011 [PubMed: 23272990]
- Cella D, Lai JS, Stone A. Self-reported fatigue: One dimension or more? Lessons from the functional assessment of chronic illness therapy-fatigue (FACIT-F) questionnaire. *Supportive Care in Cancer*. 2011; 19(9):1441–1450. DOI: 10.1007/s00520-010-0971-1 [PubMed: 20706850]
- Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. *Journal of Rheumatology*. 2005; 32(5):811–819. [PubMed: 15868614]
- Dapuerto JJ, Del Carmen Abreu M, Francolino C, Levin R. Psychometric assessment of the MSAS-SF and the FACIT-fatigue scale in Spanish-speaking patients with cancer in Uruguay. *Journal of Pain and Symptom Management*. 2014; 47(5):936–945. DOI: 10.1016/j.jpainsymman.2013.06.020 [PubMed: 24291293]
- Deboeck PR. Estimating Dynamical Systems: Derivative Estimation Hints from Sir Ronald A. Fisher. *Multivariate Behavioral Research*. 2010; 45(4):725–745. [PubMed: 26735716]
- Estabrook R. Evaluating Measurement of Dynamic Constructs: Defining a Measurement Model of Derivatives. *Psychological Methods*. 2015; 20(1):117. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24364383>. [PubMed: 24364383]
- Haas NB, Manola J, Ky B, Flaherty KT, Uzzo RG, Kane CJ, ... DiPaola RS. Effects of adjuvant sorafenib and sunitinib on cardiac function in renal cell carcinoma patients without overt metastases: Results from ASSURE, ECOG 2805. *Clinical Cancer Research*. 2015; 21(18):4048–4054. DOI: 10.1158/1078-0432.CCR-15-0215 [PubMed: 25967143]
- Haas NB, Manola J, Uzzo RG, Flaherty KT, Wood CG, Kane C, ... Dipaola RS. Adjuvant sunitinib or sorafenib for high-risk, non-metastatic renal-cell carcinoma (ECOG-ACRIN E2805): A double-blind, placebo-controlled, randomised, phase 3 trial. *The Lancet*. 2016; 387(10032):2008–2016. DOI: 10.1016/S0140-6736(16)00559-6
- Hagell P, Höglund A, Reimer J, Eriksson B, Knutsson I, Widner H, Cella D. Measuring Fatigue in Parkinson's Disease: A Psychometric Study of Two Brief Generic Fatigue Questionnaires. *Journal of Pain and Symptom Management*. 2006; 32(5):420–432. DOI: 10.1016/j.jpainsymman.2006.05.021 [PubMed: 17085268]
- Horn JL. A rationale and test for the number of factors in factor analysis. *Psychometrika*. 1965; 30:179–185. [PubMed: 14306381]
- Jabrayilov R, Emons WHM, de Jong K, Sijtsma K. Longitudinal measurement invariance of the Dutch Outcome Questionnaire-45 in a clinical sample. *Quality of Life Research*. 2017
- Kaiser HF. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*. 1960; 20:141–151. Retrieved from <http://www.garfield.library.upenn.edu/classics1986/A1986E107600001.pdf>. DOI: 10.1177/001316446002000116
- Kenny DA, Kaniskan B, McCoach DB. The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*. 2015; 44(3):486–507. DOI: 10.1177/0049124114543236
- Kline, RB. Principles and practice of structural equation modeling. 2005.
- Kosinski M, Gajria K, Fernandes A, Cella D. Qualitative validation of the FACIT-Fatigue scale in systemic lupus erythematosus. *Lupus*. 2013; 22(5):422–430. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23423250>. DOI: 10.1177/0961203313476360 [PubMed: 23423250]
- Kwakkenbos L, Willems LM, Baron M, Hudson M, Cella D, Van Den Ende CHM, Thombs BD. The comparability of English, French and Dutch scores on the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F): An Assessment of differential item functioning in patients with systemic sclerosis. *PLoS ONE*. 2014; 9(3):doi: 10.1371/journal.pone.0091979
- McArdle, JJ. Dynamic but structural equation modeling of repeated measures data. In: Nesselroade, JR., Cattell, RB., editors. *Handbook of multivariate experimental psychology*. 2. New York: Plenum Press; 1988. p. 561–614.
- Meredith W. Notes on factorial invariance. *Psychometrika*. 1964a; 29:177–185.
- Meredith W. Rotation to achieve factorial invariance. *Psychometrika*. 1964b; 29:186–206.

- Meredith W. Measurement invariance, factor analysis and factor invariance. *Psychometrika*. 1993; 58:525–543.
- Molenaar PCM. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology – This time forever. *Measurement: Interdisciplinary Research and Perspectives*. 2004; 2:201–218.
- Neale, MC., Hunter, MD., Pritikin, JN., Zahery, M., Brick, TR., Kirkpatrick, RM., ... Boker, SM. *OpenMx 2.0: Extended Structural Equation and Statistical Modeling*. Springer; New York LLC: 2015.
- Oort FJ. Using structural equation modeling to detect response shifts and true change. *Qual Life Res*. 2005; 14(3):587–598. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16022054>. [PubMed: 16022054]
- Pentz MA, Chou C-P. Measurement Invariance in Longitudinal Clinical Research Assuming Change from Development and Intervention. *Journal of Consulting and Clinical Psychology*. 1994; 62(3): 450–462. DOI: 10.1037/0022-006X.62.3.450 [PubMed: 8063972]
- Ramsay, J., Hooker, G., Graves, S. *Functional Data Analysis*. 2009. Retrieved from <http://www.springerlink.com/index/10.1007/978-0-387-98185-7>
- Widaman KF, Ferrer E, Conger RD. Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*. 2010; 4(1): 10–18. DOI: 10.1111/j.1750-8606.2009.00110.x [PubMed: 20369028]
- Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *Journal of Pain and Symptom Management*. 1997; 13(2):63–74. DOI: 10.1016/S0885-3924(96)00274-6 [PubMed: 9095563]

Figure 1.
Scree Plots for Raw and Derivative-Transformed Data

Figure 2.

Comparing Zeroth and Higher Derivative Loadings

Note. Plot of zeroth and first derivative loadings from MMOD. Heavier line indicates 1:1 relationship, while lighter line indicates line of best fit between two sets of loadings.

Table 1

Single Occasion One Factor Models

Model	χ^2	df	p	CFI	TLI	RMSEA
T1	.866	.839	.134			
T2	.886	.863	.143			
T3	.888	.866	.140			
T1 w/Residual Correlation	93.298	1	< .001	.891	.867	.122
T2 w/Residual Correlation	81.484	1	< .001	.902	.881	.134
T3 w/Residual Correlation	118.530	1	< .001	.913	.894	.124

Note. Model fit statistics for single factor models fit to each of the three timepoints of FACIT-Fatigue data individually. Models with residual correlation include a path between items AN5 and AN7, which are the only two items in the scale phrased positively, with higher endorsement indicated lack of fatigue. χ^2 tests reflect differences between nested models at each timepoint to test for significance of AN5-AN7 residual correlation (e.g., T1 vs T1 w/Residual Correlation).

Table 2

Longitudinal & MMOD Invariance Tests

Model	χ^2	df	p	CFI	TLI	RMSEA
Longitudinal Free				.892	.878	.076
Longitudinal Invariant	29.608	24	0.198	.892	.882	.075
MMOD Free				.871	.854	.075
MMOD Invariant	64.999	24	< .001	.867	.856	.074
MMOD Free				.871	.854	.075
MMOD Zeroth vs. Change	15.402	12	0.220	.871	.857	.074
MMOD Invariant	49.597	12	< .001	.867	.856	.074

Note. Fit statistics for longitudinal and MMOD models of FACIT-Fatigue. All models include a residual correlation between AN5 & AN7 and longitudinal autocorrelations for each item. χ^2 tests reflect differences between nested free and invariant models to test for invariance.

Table 3

Cross-Sectional and MMOD Factor Loadings

Code	Item	T1	Zeroth	First	Second
HI7	I feel fatigued	0.828	0.889	0.775	0.748
HI12	I feel weak all over	0.815	0.863	0.762	0.753
AN1	I feel listless (washed out)	0.825	0.885	0.779	0.773
AN2	I feel tired	0.811	0.875	0.804	0.776
AN3	I have trouble starting things because I am tired	0.837	0.913	0.818	0.819
AN4	I have trouble finishing things because I am tired	0.823	0.895	0.799	0.776
AN5	I have energy	-0.597	-0.743	-0.504	-0.439
AN7	I am able to do my usual activities.	-0.544	-0.683	-0.386	-0.343
AN8	I need to sleep during the day	0.505	0.646	0.503	0.501
ANI2	I am too tired to eat	0.473	0.662	0.529	0.504
ANI4	I need help doing my usual activities	0.565	0.733	0.516	0.510
ANI5	I am frustrated by being too tired to do the things I want to do	0.808	0.876	0.763	0.747
ANI6	I have to limit my social activity because I am tired.	0.796	0.878	0.763	0.721

Note. Standardized factor loadings for cross sectional baseline model and MMOD analyses of derivative-transformed longitudinal data.