Annals of Depression and Anxiety

Open Access

Austin Publishing Group

Research Article

A Comparison of Individual Change using Item Response Theory and Sum Scoring on the Patient Health Questionnaire-9: Implications for Measurement-Based Care

Jones SMW[1]*, Crane PK[2] and Simon G[3]

[1]Fred Hutchinson Cancer Research Center, US

[2]Department of Medicine, University of Washington, US

[3]Senior Investigator, Kaiser Permanente Washington Health Research Institute, US

***Corresponding author:** Jones SMW, Assistant Member, Fred Hutchinson Cancer Research Center, Fairview Ave, Seattle, WA 98109, US

Received: January 22, 2019; **Accepted:** March 25, 2019; **Published:** April 01, 2019

Abstract

We examined change over time in depression with standard sum *vs*. Item Response Theory (IRT) scoring. Patient Health Questionnaire 9 item responses were extracted from the electronic health records of 5,405 people receiving depression treatment at the start of treatment and 30 to 180 days later. We used four methods to classify change: the Reliable Change Index (RCI), the 5-point change and 50% change from baseline for sum scores and the z-test for IRT scoring. The 5-point change and 50% change from baseline are both Health Effectiveness Data and Information Set measures. The z-test mostly agreed with the RCI, 5-point change or 50% change. More people had change using 5-point change or 50% change but not IRT scoring than no change using 5-point or 50% change but change using IRT scoring. Kappas between changes on IRT and sum scores ranged from 0.620 to 0.813. This difference in agreement is likely meaningful at the individual, patient level. People classified differently between IRT and sum scoring had moderate symptom change. Differences in conclusions from IRT and sum scoring may be most relevant in challenging clinical situations such as small or moderate symptom change.

Keywords: Depression; Treatment response; Item response theory; Change scores

Introduction

Item Response Theory models (IRT) have been increasingly used as an alternative to classical test theory in measure development and validation for psychiatric outcomes such as depression and anxiety [1]. IRT scoring may have more precision in distinguishing statistically significant individual differences in change over time [2]. A cross-sectional study found that even among people with the same standard sum score, IRT scores were associated with external criteria in the hypothesized direction [3], suggesting that IRT scoring may be more informative of actual level of depression or other symptoms in treatment compared to standard sum scores. Simulation studies demonstrate that IRT scoring may reduce bias in estimating rates of change over time compared to standard sum scoring [4]. Part of this reduction in bias may stem from IRT models not assuming that error is constant along the continuum of a measure, unlike classical test theory [2]. Although IRT scores and sum scores are highly correlated, even the small amount of disagreement between the scores may have impact at the individual patient level [5].

While there may be some psychometric advantages of IRT compared to classical test theory, different scoring methods may have different usefulness in measurement-based care. Measurement-based care is the use of patient-reported data in healthcare treatment, Primarily Patient-Reported Outcomes (PROs) [6-9], adoption in community is variable and below 20% [10]. New Health Effectiveness Data and Information Set (HEDIS) quality metrics emphasizing measurement based care [11] are expected to accelerate use of measurement-based care, assessing individual change in measurement-based care is particularly difficult and remains a barrier to implementation [5,10]. IRT may be one way to address this challenge. But the benefits of IRT scoring in measurement-based care needs to be considered against the practical advantages of standard sum scoring (simpler, easier, more transparent to clinicians). For example, nearly half of practicing clinical psychologists are in private practices [12] and only 15% of psychiatric hospitals have electronic medical records [13]. Implementing a complicated scoring system like IRT would be challenging in these settings as they do not have the infrastructure of large medical-surgical hospitals or academic centers. Research on different scoring methods for individual change have been mixed [14-18].

The aim of this study was to compare agreement between IRT scoring to standard sum scale scoring in classifying change from depression treatment initiation to follow-up on the Patient Health Questionnaire-9 (PHQ9). HEDIS focuses on simple, easy to compute measures of change such as 50% change from baseline [11] and more sophisticated measures of statistically significant change such as from IRT would only be needed if these methods disagreed substantially and were not interchangeable. Measurement-based care includes evaluating whether the initial treatment choice was successful, so determining change of symptoms from treatment initiation could help inform clinical decision making. We therefore focused on

Citation: Jones SMW, Crane PK and Simon G. A Comparison of Individual Change using Item Response Theory and Sum Scoring on the Patient Health Questionnaire-9: Implications for Measurement-Based Care. Ann Depress Anxiety. 2019; 6(1): 1098.

whether IRT and standard scoring provided different results on whether the initial treatment was effective or not. We also specified two sets of change measures for sum scores, statistically significant change by the Reliable Change Index [19] and general guidelines [20], for comparison to IRT scoring.

Methods

Population and procedures

Data were collected from the Electronic Health Records (EHR) of people starting treatment for depression (psychotherapy or antidepressants) in three integrated health systems: Kaiser Permanente Washington (formerly Group Health), Kaiser Permanente Colorado, and HealthPartners (n=5,420, see flow chart in Figure 1). A new episode of either antidepressant medication or psychotherapy was defined by a psychotherapy visit or a filled antidepressant prescription associated with a diagnosis of depression, preceded by at least 180 days without a psychotherapy visit or antidepressant prescription. Data were extracted for the period between January 1, 2010 and December 31, 2012. PHQ9 item responses were collected at baseline, defined as when participants were first starting depression treatment within our time window, and at a follow-up health care visit that occurred at least 30 days after baseline but no more than 180 days after baseline. Limited demographic information was collected from the EHR including age, sex, race/ethnicity and presence of medical comorbidities. As this study was a secondary analysis of data collected from another study, we did not have specific diagnoses for comorbidities nor number of comorbidities, though we had data on whether medical comorbidities were present as measured by the Charlson Comorbidity Index [21]. Responsible Institutional Review Boards for each health system reviewed all study procedures and approved a waiver of consent for use of de-identified records data for this research (IRB#213058). Study procedures complied with all ethical standards including the Helsinki Declaration.

Measures

The PHQ9 [22,23] is a questionnaire-based measure of depressive symptoms based on the Diagnostic and Statistical Manual [24,25]. It has nine items, each representing one of the symptoms assessed for a diagnosis of depression: depressed mood, anhedonia, sleep disturbance, fatigue, appetite or weight changes, feeling guilty, perceived cognitive changes, psychomotor agitation or retardation, and suicidal ideation. Response categories for each item refer to how frequently the respondent experienced each of these symptoms in the past two weeks. Respondents rate their symptoms on a scale of 0 (not at all), 1 (several days), 2 (more than half the days), or 3 (nearly every day). Traditional sum scores range from 0 to 27.

Statistical Analyses

PHQ9 scoring

The PHQ9 has a single factor structure [26-29] and has been found to be sufficiently unidimensional for IRT analyses [30]. An exploratory factor analysis using maximum likelihood estimation from the complete cases (n=5,351) in our follow-up sample showed a unidimensional structure (first eigenvalue=4.166, second eigenvalue=0.982). We used the graded response model [31] to determine IRT parameters for the PHQ9 using the follow-up sample (n=5,420); the resulting scores had a mean of 0 and SD of 1 (z-scores).

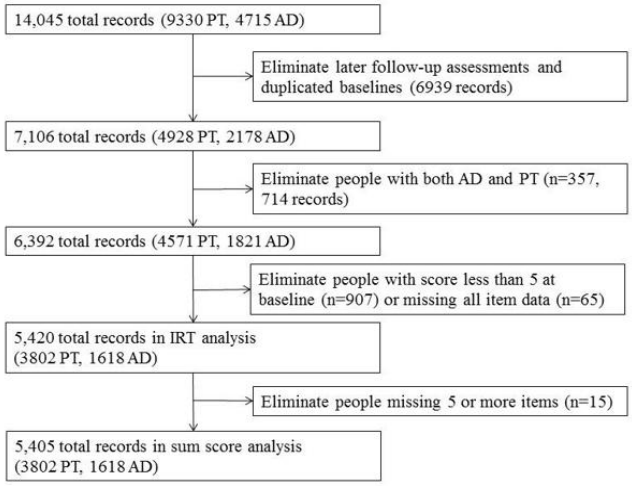

Figure 1: Flow chart of participants included and excluded. PT: Psychotherapy; AD: Antidepressants; PHQ9: Patient Health Questionnaire 9.

The graded response model estimates two types of parameters per item. First is the slope parameter, which indicates how accurately the item reflects the underlying construct (depression in this case). The second parameter type is the severity or level parameter that indicates how much of the underlying construct a person has to have before they respond yes to a particular item, or in the case of multi-category responses, to that particular response category. These parameters can then be used to estimate how much information the measure provides along the continuum of the construct. We chose the follow-up sample to ensure broadest spread of depression levels; since we studied initiation of depression treatment, most patients were expected to be experiencing significant depressive symptoms at baseline. As depression improves for some patients during treatment but not others, this means the follow-up scores would have better range of the construct for estimating item parameters. As our sample came from integrated health systems that tend to have patients slightly different from the general population, we elected to create item parameters rather than use already published parameters [32]. We used IRTPRO 2.0 (Skokie, IL, Scientific Software International) to calculate item parameters and scores at the follow-up visit. We then used the resulting item parameters to calculate IRT scores for the baseline data using IRTPRO and Expectation A Posteriori (EAP) estimation. IRT scores account for different levels of error for more extreme (very high or very low) scores and throughout the continuum of scores whereas sum scores treat all scores as equally reliable.

For standard scoring, we used the traditional method of summing the items. A very small number of people (n=15, 0.3%) were missing more than 50% of the items at one of the assessments and we elected to exclude these individuals from the standard scoring analyses. We choose 50% as this is what is currently used in the Patient-Reported Outcomes Measurement Information System (PROMIS) measures [33]. For the people included in the standard sum scoring (n=5,405), we used individual mean imputation when there were fewer than 5 items missing. This was calculated by using the mean of the items the person did answer to impute the missing items. This did not use the mean for the sample but rather the mean for items that the individual answered.

Table 1: Descriptive statistics for the total sample. Sum scores refer to the scores calculated from the items reported in the health record. Values are means (standard deviations) and percentages for column (n) unless otherwise noted. Concordance was determined using IRT scores (z-tests) and the Reliable Change Index (RCI) for sum scores.

	Concordant IRT and Sum Score Results	Discordant IRT and Sum Score Results	Total sample
	N=4981	N=424	N=5405
Sum score, baseline	14.0 (5.6)	13.8 (5.6)	14.0 (5.6)
Sum score, follow-up	12.4 (6.3)	12.1 (6.1)	12.3 (6.3)
IRT score, baseline	0.27 (0.81)	0.23 (0.79)	0.27 (0.81)
IRT score, follow-up	0.01 (0.93)	-0.07 (0.93)	0.00 (0.93)
Time to follow-up (days)	84.7 (43.9)	85.6 (43.7)	84.8 (43.9)
Time to follow-up (days), median (interquartile range)	71 (48, 119)	74 (48, 121.25)	71 (48, 119)
Treatment			
Medication	30% (1479)	29% (124)	30% (1603)
Therapy	70% (3502)	71% (300)	70% (3802)
% female	72% (3565)	71% (300)	72% (3865)
% with comorbidity	17% (853)	17% (72)	17% (925)
Age	46.2 (16.8)	46.8 (16.6)	46.2 (16.8)
Race/Ethnicity			
White	78% (3900)	76% (323)	78% (4223)
Black/AA	6% (280)	6% (27)	6% (307)
Asian	3% (159)	3% (12)	3% (171)
Hispanic	5% (268)	7% (28)	5% (296)
Other	8% (374)	8% (34)	8% (408)

AA: African American; IRT: Item Response Theory. Percentages may not add up to 100% due to rounding. Comorbidity is based on the Charlson Comorbidity Index.

Statistical Analysis

Assessment of change

We used four approaches to categorize people as having meaningful improvement, meaningful worsening or no meaningful change in depressive symptoms, three appropriate to standard sum scoring and the other appropriate for IRT scoring. We used the Reliable Change Index (RCI) [19] for statistically significant change in sum scores. We used the previously published Minimal Clinically Important Difference (MCID) of 5 points on the PHQ9 [34] and 50% change from baseline [11] for clinically meaningful change in standard sum scoring. The z-test [35] showed statistically significant change for IRT scoring.

$$Ztest = \left(\theta baseline - \theta\, followup\right) / \left(\sqrt{\sigma^2 baseline + \sigma^2\, followup}\right)$$

An alpha level of 0.05 was used for determining significant improvement at the individual level for both the z-test and RCI.

The RCI is calculated from the difference between two scores (baseline and follow-up in this case) and the difference is then divided by the sample standard error of the difference between the two scores, calculated from the standard deviation of the sample and the reliability of the scores. In the equation below, the X's represent the two scores, SD is the sample standard deviation (baseline SD in this study) and r is the reliability of the PHQ9.

$$RCI = \left(Xpre - Xpost\right) / \left(\sqrt{2} * \left(SD * \sqrt{1-r}\right)\right)$$

Participants were classified as improved by the RCI if the RCI showed statistically significant decreases on the PHQ9 and as worsened by the RCI if the RCI showed their PHQ9 scores had significantly increased. All other patients were classified as no change.

The 5-point change (MCID) for the PHQ9 was established as two standard errors of measurement from a group of people who had completed depression treatment in a different study [34]. Unlike the RCI, since the 5-point changed is already established and requires no calculations, it is readily available at the time of visits. The 50% change in symptoms from baseline is a standard measure of clinical response in the treatment of depression [36]. For the 5-point change, cases were classified as improved if their symptoms decreased by 5 or more points and worsened if their symptoms increased by 5 or more points. For the 50% change, cases were classified as improved if their symptoms decreased by 50% or more of the baseline level (i.e. if a person with a score of 20 at baseline had a score of 10 or lower at follow-up) and worsened if symptoms increased by 50% or more of the baseline level.

The z-test is calculated similarly to the RCI in that it involves the difference between two IRT scores, but unlike the RCI the difference is divided by a pooled standard error of measurement derived from the standard error of measurement associated with IRT-based scores. Standard IRT software produces not only an estimate of each person's score, but also an estimate of the standard error of measurement, indicating how precise the score is. Like the RCI, significant decreases in the PHQ9 by the z-test were classified as improved, significant increases were classified as decreased and all others were classified as no change.

We used SPSS 22 (IBM Corp., Armonk, NY) and SAS 9.3 (SAS

Table 2: Percentage of the sample in each group as categorized by IRT (z-test) and sum scores (RCI, 5-point change, 50% change) and PHQ9 scores. Negative numbers for change indicate symptoms worsened. Note when methods disagreed, the method not identified did not find any meaningful change. For example, **'Methods Disagreed:** IRT found improvement' means those patients were classified as improved by IRT scoring but as no change by the identified sum score method. Methods Agreed means the IRT scores and the identified sum score change measure classified the patient similarly (either improved, worsened or no change).

Sum score measure of change	Group	IRT change measured through z-test	Mean, baseline PHQ9	Mean, follow-up PHQ9	PHQ9 change (baseline minus follow-up)
RCI	Methods Agreed: Improved	18.6	17.9	6.6	11.28
	Methods Agreed: Worsened	8.3	9.25	19.46	-10.18
	Methods Agreed: No Change	65.2	13.47	13.1	0.37
	Methods Disagreed: IRT found improvement	2.7	14.24	8.73	5.54
	Methods Disagreed: Sum scores found improvement	2.4	17.31	9.78	7.49
	Methods Disagreed: IRT found worsening	1.2	11.71	17.21	-5.5
	Methods Disagreed: Sum scores found worsening	1.6	9.52	17.02	-7.5
	Total percent different between IRT & sum scores	7.8			
5-point change	Methods Agreed: Improved	21	17.49	6.84	10.64
	Methods Agreed: Worsened	9.4	9.51	19.17	-9.63
	Methods Agreed: No Change	53.1	13.42	13.14	0.29
	Methods Disagreed: IRT found improvement	0.2	13.46	9.54	3.92
	Methods Disagreed: Sum scores found improvement	9.3	16.06	10.11	5.92
	Methods Disagreed: IRT found worsening	0.1	15.75	19.25	-3.5
	Methods Disagreed: Sum scores found worsening	6.8	10.71	16.64	-5.92
	Total percent different between IRT & sum scores	16.4			
50% Change	Methods Agreed: Improved	15.8	16.49	4.86	11.61
	Methods Agreed: Worsened	8.7	9.01	18.95	-9.92
	Methods Agreed: No Change	59.1	14.44	13.68	0.75
	Methods Disagreed: IRT found improvement	5.5	20.18	12.64	7.55
	Methods Disagreed: Sum scores found improvement	3.9	8.89	3.71	5.16
	Methods Disagreed: IRT found worsening	0.8	15.79	21.57	-5.79
	Methods Disagreed: Sum scores found worsening	6.3	7.66	13.12	-5.45
	Total percent different between IRT & sum scores	16.4			

IRT: Item Response Theory; RCI: Reliable Change Index; PHQ: Patient Health Questionnaire

Institute, Cary, NC) to calculate descriptive statistics and compare change over time. Participants could be categorized into one of nine categories, although two categories did not occur in this sample (see Figure 1 in supplementary materials). Agreement between IRT and sum score change classification was examined using the kappa statistic and percent agreement. We also compared agreement between the different sum score methods. Participants in groups one through three were classified as concordant and participants in groups four through seven were classified as discordant.

Results

Demographic and clinical characteristics for the sample are reported in Table 1. Most were middle-aged (mean 46.2 years old) and most were female (72%). More people were starting treatment with psychotherapy (70%) compared to medication (30%).

Item response theory model

The slope parameters from the graded response model for all items were in a similar range to those previously reported for the PHQ9 (See Supplementary Materials, Table 1; ranging from 1.28 to 2.90, [32]). The IRT model had an RMSEA of 0.05. The standard error curve for the PHQ9 (see Supplementary Materials, Figure 2) indicated that depression at the follow-up visit was measured with the least amount of error between 1.5 standard deviations below the mean to 1.5 standard deviations above the mean. This suggests that measurement error was highest for those with no measurable depressive symptoms to very low depressive symptoms or those with extremely high levels of depression. For the whole sample, IRT scores decreased between baseline and follow-up by an average of 0.27 points (CI: 0.25, 0.30, $t(5419)=20.42$, $p<0.001$) and standard sum scores decreased by an average of 1.64 points (CI: 1.47, 1.82, $t(5404)=18.41$, $p<0.001$).

Change for IRT scoring *vs*. standard scoring

Using the RCI (sum scores) and z-test (IRT scores), most participants were classified similarly but a small percentage were

classified differently (see Table 2). Most participants (65.2%) were classified as having no change by both IRT and sum scoring. Several participants' symptoms were classified as improved by both methods (18.6%) or worsened by both methods (8.3%). A small number of people had their symptoms classified as not significantly changed by IRT scoring but as significantly worse (1.6%) or improved (2.4%) by the RCI. For example, one person had a baseline score of 17 and follow-up score of 10. This person endorsed all symptoms at baseline, except psychomotor agitation/retardation. At the follow-up, this person still endorsed all symptoms except psychomotor agitation/retardation and appetite changes, and most symptoms improved slightly but their fatigue worsened. The change in overall sum score was enough to be significant by the RCI but not the z-test. Symptoms for a similar number of people were classified as not significantly changed by standard sum scoring but significantly worse (1.2%) or improved (2.6%) by IRT scoring. For example, one person had a baseline score of 20 and a follow-up of 25, endorsing all the symptoms except suicidal ideation at baseline. This person did endorse all symptoms including suicidal ideation at follow-up so although the RCI did not take the unique implications of suicidal ideation into account, the IRT scoring and z-test did show this as a worsening of symptoms. In total, 92.2% of people were classified similarly but symptoms for 7.8% were classified differently between sum scoring using the RCI and IRT scoring with the z-test. The kappa for agreement between the z-test and the RCI was 0.813.

We observed greater discrepancies when comparing the 5-point change approach for standard scoring to the z-test for IRT scoring. Most participants were classified as having no change by both IRT and the 5-point change approach (53.1%), with some classified as either worsening symptoms by both methods (9.4%) or improving by both methods (21.0%). However, among those classified as no change by the 5-point change approach, few were classified as improved (0.2%) or worse (0.1%) by IRT scoring. Among the whole sample, substantial numbers were classified as either worsening symptoms (6.8%) or improving symptoms using the 5-point change approach (9.3%) but not by IRT scoring. For example, one individual had a baseline score of 21 and a follow-up score of 16. This individual endorsed all symptoms except suicidal ideation at both times, however, certain symptoms (anhedonia, appetite, guilt and cognition) reduced by one or two points leading to the improvement by the 5-point change approach but not by the z-test. Overall, 83.6% of people were classified similarly by the 5-point change approach and IRT scores, but 16.4% were classified differently. The kappa between z-test groups and the 5-point change approach groups was 0.644.

Like the 5-point change approach, a greater discrepancy was found between the z-test and 50% change than between the RCI and z-test. Over half the sample (59.1%) was classified as no change by both IRT scores and 50% change while 15.8% were classified as improved and 8.7% were classified as worsened by both methods. A very small percentage of the sample was classified as worse by IRT with no difference by 50% change (0.8%). More substantial numbers were classified as no change by IRT but improved (3.9%) or worsened (6.3%) by 50% change. Another 5.5% were classified as improved by IRT scores but not by the 50% change. Overall, 16.4% of the sample differed between classifications from the IRT scores/z-test and 50% change. The kappa between z-test classifications and 50% change classifications was 0.620.

Table 3: Comparison of RCI with 5-point change and 50% change classifications from post-hoc analyses. Note when methods disagreed, the method not identified did not find any meaningful change. For example, **'Methods Disagreed:** RCI found improvement' means those patients were classified as improved by RCI scoring but as no change by the other sum score method identified in the column header. Methods Agreed means the identified sum score change methods classified the patient similarly (either improved, worsened or no change).

	Percent of total sample	
Group	5-point change	50% change
Methods Agreed: Improved	21.2	15.6
Methods Agreed: Worsened	9.9	9.6
Methods Agreed: No Change	53.3	59.7
Methods Disagreed: RCI found improvement	0	5.4
Methods Disagreed: Other method found improvement	9.3	4.1
Methods Disagreed: RCI found worsening	0	0.2
Methods Disagreed: Other method found worsening	6.3	5.4
Total percent different	15.6	15.1

When examining the baseline and follow-up means of the PHQ9 sum scores for the different groups (see Table 2), IRT and sum scoring differed in classifying change between one standard deviation below the mean to one standard deviation above the mean (baseline sum scores between 8.2 and 19.4 for the four discordant groups). For the groups in which IRT and sum scores agreed, large average changes of 10 points were seen for the two groups reporting improvement or worsening of symptoms and practically no change on the means for the groups with no change by either method. However, for the groups in which IRT scores and sum scores by any method disagreed, mean differences from baseline to follow-up were smaller ranging from 3.5 to 7.6 points.

Post-hoc analyses: change for different sum scoring methods

Because the proportion of people showing change by the 5-point change approach but not IRT scoring was so high compared to the converse (change by IRT but not the 5-point change approach), we conducted post-hoc analyses comparing the classifications of the RCI and the 5-point change approach to determine whether these results were due to the 5-point change approach (see Table 3). We also conducted analyses comparing the RCI with the 50% change criterion. There were no cases in which the 5-point approach suggested no change but the RCI did suggest change. However, a large minority of cases were classified as worse (6.8%) or improved (9.3%) by the 5-point change approach but not the RCI. For 50% change to RCI comparison, results were like the IRT and 50% change results. Few cases showed worsening by the RCI but no difference by 50% change (0.2%) but several cases showed improvement by the RCI but not 50% change (5.4%) as well as no change by the RCI but improvement (4.1%) and worsening by 50% change (5.4%).

Discussion

This study compared IRT and sum score-based change from a commonly-used depression measure among people initiating antidepressant or psychotherapy treatment for depression. We considered the clinical implications of a variety of scoring methods

for determining meaningful change. When comparing statistically significant change using the RCI (standard sum scoring) and the z test for IRT scoring, most people's depressive symptoms were classified the same way by both methods. Similar proportions of individuals were classified with symptom change by the RCI but not z-test and with symptom change by the z-test but not the RCI. When comparing meaningful change using the 5-point change approach or 50% change from baseline with standard sum scoring and the z test for IRT scoring, over 16% of the sample was classified differently. The 5-point change approach tended to suggest improvement or worsening of symptoms whereas IRT scoring indicated no actual change. Our results suggest that different scoring methods may affect measures showing change as well as tests suggesting no change. Results also showed that differences in meaningful change between IRT and sum scores occurred at moderate symptom change levels. A large portion of our sample did not improve, consistent with other studies of real-world depression treatment given the length of follow-up [37].

The means of the sum scores and change in sum scores suggest that different scoring methods may provide different information in areas that would be of great importance to clinicians using measurement-based care. The disagreements also occurred with average absolute change values ranging from 3.5 to 7.6 points where it may be challenging to tell if change has occurred. A 10-point or more change on the 0-27 point range of the PHQ9 is clearly a marked change and detectable by any method., A smaller changes like 3 points is less clear as either error or a true change and IRT scoring could help clinicians determine the clinical relevance of that change and whether treatment should be adjusted or continued.

The present study adds to previous research showing the benefit of using IRT scores over sum scores in the measurement of symptom change in depression treatment [15-17] by showing IRT scores are possibly providing different information when most needed by clinicians within measurement-based care [6]. At the individual level, even small amounts of disagreement such as reliability less than 0.90 can lead to incorrect significance tests [5]. Even the small amount of discordance between sum scores and IRT scores observed here suggests sum scores might not be an acceptable approximation of the true score when measuring change at the individual level, particularly as previous studies suggested IRT scores more accurately measure change [17].

One unexpected finding based on previous literature [15] was that IRT scoring showed no change when standard sum scoring showed change and this happened nearly as often as IRT scoring showing change when sum scoring did not. This possibly happened because IRT scores weight items differently whereas sum scores treat all item-level changes equally. For example, IRT scores would weight minor increases in severe symptoms (anhedonia, suicidality) more heavily than minor decreases in less severe symptoms (sleep disturbance, fatigue). Sum scores would not account for these differences and, if the decrease in less severe symptoms was high enough, would lead to change by sum scores but not IRT scores. This is particularly important for measurement-based care of depression as knowing when a treatment is not working, and hence, needs to be changed, is just as important as knowing when a treatment works and should be continued.

Our results should be considered within the strengths and limitations of the study. The data came from actual clinical use of the PHQ9, increasing external validity. Also, the same items were used for both standard sum scoring and the IRT analyses so any differences in classifying symptoms are due to the scoring methods and not to the specific items used. Although the three study sites increased generalizability, the sites may differ from other clinical settings particularly given that nearly all participants had insurance and all sites were in the United States. We also created item parameters from our sample instead of using published parameters [32]. The sample was also predominantly female consistent with depression being more prevalent in women [38,39]. Our sample included more people starting psychotherapy than medication, although the difference would be expected given that the patients had to have at least one follow-up assessment. We also only considered the most recent follow-up symptom assessment. Although we had to exclude patients starting both psychotherapy and medications, it is important to note that this was only 5% of the sample likely because patients will often try one treatment for a period of time and then add another treatment if the first treatment does not bring enough relief.

Conclusion

This study suggests that use of IRT scores instead of sum scoring may provide different information both in when a treatment is working and when it is not working. This could affect treatment delivery within measurement-based care as clinicians would adapt or maintain a treatment approach based on whether there is meaningful change or no change. Although more research is needed to definitively determine whether the use of IRT scoring instead of standard scoring in measurement-based care affects outcomes, our results support the use of IRT scoring instead of traditional sum scoring in monitoring depression.

References

1. Pilkonis PA, Yu L, Dodds NE, Johnston KL, Maihoefer CC, Lawrence SM. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study. J Psychiatr Res. 2014; 56: 112-119.

2. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. J Pers Assess. 2005; 84: 228-238.

3. Snitz BE, Yu L, Crane PK, Chang CC, Hughes TF, Ganguli M. Subjective cognitive complaints of older adults at the population level: an item response theory analysis. Alzheimer disease and associated disorders. 2012; 26: 344-351.

4. Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol. 2008; 61: 1018e8-1027e9.

5. Donaldson G. Patient-reported outcomes and the mandate of measurement. Qual Life Res. 2008; 17: 1303-1313.

6. Scott K, Lewis CC. Using measurement-based care to enhance any treatment. Cognitive and Behavioral Practice. 2015; 22: 49-59.

7. Rush AJ. Isn't It About Time to Employ Measurement-Based Care in Practice? Am J Psychiatry. 2015; 172: 934-936.

8. Guo T, Xiang YT, Xiao L, Hu CQ, Chiu HF, Ungvari GS, et al. Measurement-Based Care *Versus* Standard Care for Major Depression: A Randomized Controlled Trial With Blind Raters. Am J Psychiatry. 2015; 172: 1004-1013.

9. Fortney JC, Unützer J, Wrenn G, Pyne JM, Smith GR, Schoenbaum M, et al. A Tipping Point for Measurement-Based Care. Psychiatr Serv. 2017; 68:

179-188.

10. Fortney J, Sladek R, Unützer J, Kennedy P, Harbin H, Emmet B, et al. Fixing Behavioral Health Care in America: A National Call for Measurement-Based Care in the Delivery of Behavioral Health Services. The Kennedy Forum. 2015.

11. Health Effectiveness Data and Information Set 2018 Measures. National Committee for Quality Assurance, Washington, DC: National Committee for Quality Assurance. 2017.

12. Hamp A, Stamm K, Lin L, Christidis P. 2015 APA Survey of Psychology Health Service Providers. American Psychological Association, Studies ACfW. 2016.

13. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. ONC Data Brief, no. 35. . Washington DC: Office of the National Coordinator for Health Information Technology. 2016.

14. Zimmerman M, D'Avanzato C, Attiullah N, Friedman M, Toba C, Boerescu DA. Scoring rules and rating formats of Self-report Depression Questionnaires: a comparison of approaches. Psychiatry Res. 2014; 218: 225-228.

15. Brouwer D, Meijer RR, Zevalkink J. Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. Psychotherapy research: journal of the Society for Psychotherapy Research. 2013; 23: 489-501.

16. Jabrayilov R, Emons WH, Sijtsma K. Comparison of classical test theory and item response theory in individual change assessment. Applied Psychological Measurement. 2016; 40: 559-572.

17. Gorter R, Fox JP, Twisk JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Med Res Methodol. 2015; 15: 55.

18. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989; 10: 407-415.

19. Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology. 1991; 59: 12-19.

20. Cook CE. Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. J Man Manip Ther. 2008; 16: E82-E83.

21. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. J Clin Epidemiol. 1994; 47: 1245-1251.

22. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16: 606-613.

23. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA. 1999; 282: 1737-1744.

24. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV-TR. 4th ed., text revision ed. Washington, DC: American Psychiatric Association; 2000.

25. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Arlington, VA: American Psychiatric Publishing; 2014.

26. Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract. 2008; 58: 32-36.

27. Dum M, Pickren J, Sobell LC, Sobell MB. Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. Addict Behav. 2008; 33: 381-387.

28. Hansson M, Chotai J, Nordstöm A, Bodlund O. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. Br J Gen Pract. 2009; 59: e283-e288.

29. Kalpakjian CZ, Toussaint LL, Albright KJ, Bombardier CH, Krause JK, Tate DG. Patient health Questionnaire-9 in spinal cord injury: an examination of factor structure as related to gender. J Spinal Cord Med. 2009; 32: 147-156.

30. Crane PK, Gibbons LE, Willig JH, Mugavero MJ, Lawrence ST, Schumacher JE, et al. Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). AIDS Care. 2010; 22: 874-885.

31. Samejima F. Estimation of a latent ability using a response pattern of graded scores. 1969; 34.

32. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess. 2014; 26: 513-527.

33. PROMIS Short Forms for Open Distribution User Instructions. National Institutes of Health; 2012.

34. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care. 2004; 42: 1194-1201.

35. Guo J, Drasgow F. Identifying cheating on unproctored Internet tests: the z-test and the likelihood ratio test. International Journal of Selection and Assessment. 2010; 18: 351-364.

36. O'Connor M, Butcher I, Hansen CH, Kleiboer A, Murray G, Sharma N, et al. Measuring improvement in depression in cancer patients: a 50% drop on the self-rated SCL-20 compared with a diagnostic interview. Gen Hosp Psychiatry. 2010; 32: 334-336.

37. Licht-Strunk E, van der Windt DA, van Marwijk HW, de Haan M, Beekman AT. The prognosis of depression in older patients in general practice and the community. A systematic review. Fam Pract. 2007; 24: 168-180.

38. Hays RD. Change in health-related quality of life over 3 months in chiropractic patients with chronic low back pain. UCLA Department of Medicine GIM/ Health Services Seminar Series; Los Angeles, CA 2018.

39. Riolo SA, Nguyen TA, Greden JF, King CA. Prevalence of depression by race/ethnicity: findings from the National Health and Nutrition Examination Survey III. Am J Public Health. 2005; 95: 998-1000.

Ann Depress Anxiety - Volume 6 Issue 1 - 2019
ISSN : 2381-8883 | www.austinpublishinggroup.com

Citation: Jones SMW, Crane PK and Simon G. A Comparison of Individual Change using Item Response Theory and Sum Scoring on the Patient Health Questionnaire-9: Implications for Measurement-Based Care. Ann Depress Anxiety. 2019; 6(1): 1098.