Item-Response-Theory Analysis of Two Scales for Self-Efficacy for Exercise Behavior in People With Arthritis

Thelma J. Mielenz, Michael C. Edwards, and Leigh F. Callahan

Benefits of physical activity for those with arthritis are clear, yet physical activity is difficult to initiate and maintain. Self-efficacy is a key modifiable psychosocial determinant of physical activity. This study examined two scales for self-efficacy for exercise behavior (SEEB) to identify their strengths and weaknesses using item response theory (IRT) from community-based randomized controlled trials of physical activity programs in adults with arthritis. The 2 SEEB scales included the 9-item scale by Resnick developed with older adults and the 5-item scale by Marcus developed with employed adults. All IRT analyses were conducted using the graded-response model. IRT assumptions were assessed using both exploratory and confirmatory factor analysis. The IRT analyses indicated that these scales are precise and reliable measures for identifying people with arthritis and low SEEB. The Resnick SEEB scale is slightly more precise at lower levels of self-efficacy in older adults with arthritis.

Keywords: physical activity, older adults, predictors

Adults with arthritis who are physically inactive tend to have more pain and disability, lower quality of life, and higher medical costs than their physically active counterparts (Kaplan, Huguet, Newsom, & McFarland, 2003). There are demographic and health-related characteristics to help identify sedentary adults who would benefit most from physical activity interventions. However, most of these factors are not modifiable and thus cannot guide interventions to increase physical activity. Numerous psychosocial and behavioral factors have been proposed to influence physical activity behavior and, given that they are modifiable, are targets of intervention (Kubiak, 2004).

Among modifiable psychological factors, self-efficacy is one of the strongest predictors of physical activity behavior (Keller, Fleury, Gregor-Holt, & Thompson, 1999; Sherwood & Jeffery, 2000). Self-efficacy significantly influences many forms of physical activity behavior, including adherence to exercise programs, maintenance of a physically active lifestyle, and sustained physical activity after

Mielenz is with the Dept. of Epidemiology, Columbia University, New York, NY. Edwards is with the Dept. of Psychology, The Ohio State University, Columbus, OH. Callahan is with Thurston Arthritis Research Center, Chapel Hill, NC.
an exercise intervention (Brassington, Atienza, Perczek, DiLorenzo, & King, 2002; Conn, 1997, 1998; DuCharme & Brawley, 1995; Garcia & King, 1991; Kubiak, 2004; McAuley, 1993; McAuley & Jacobson, 1991; McAuley, Jerome, Elavsky, Marquez, & Ramsey, 2003; Oman & King, 1998). Physical activity and exercise are key outcomes in self-efficacy-enhancing interventions such as those by Lorig, Ritter, Laurent, and Plant (2008) and other researchers (Bennett et al., 2005; Goeppinger et al., 2009; Olsen & Nesbitt, 2010) for self-management of chronic conditions such as arthritis. *Physical activity and exercise*, even though they are distinct terms, are often used interchangeably, and both will be used in this study.

The precise measurement of self-efficacy is crucial in all self-management interventions because it is such a strong determinant of exercise behavior and is modifiable. In theory, self-efficacy measures should include items reflecting a “respondent’s belief about their capacities to perform behaviors within a particular domain of functioning under circumstances that present graduations of challenge” (Forsyth & Carey, 1998, p. 561). This operational definition illustrates the difficulties of measuring the self-efficacy construct and builds a strong argument to evaluate these scales not only in the context in which they were developed and used thus far (i.e., classical test theory [CTT]) but also with state-of-the-art modern measurement theory that uses item response theory (IRT).

IRT is a collection of latent variable measurement models. Once IRT calibration has taken place it is possible to use the resulting item parameters to create scores that take into account the item-level properties. These are essentially weighted scores based on an individual’s response pattern and the item parameters. IRT scores have a number of desirable properties including an interpretable metric, increased variability, and the ability to account for differences in item-level properties (Curran, Edwards, Wirth, & Hussong, 2007; Wirth & Edwards, 2007). IRT models allow for empirically based differential weighting of items and also provide a more detailed analysis of reliability than is possible with CTT. This study examines two scales for self-efficacy for exercise behavior (SEEB).

By identifying the relative merits and weaknesses of these two scales, the precision of identifying people with arthritis and other chronic conditions and low self-efficacy will facilitate future research. The 5-item scale by Marcus (Marcus SEEB) was developed in working adults, whereas the 9-item scale by Resnick (Resnick SEEB) was specially revised for older adults (Marcus & Owen, 1992; Marcus, Selby, Niaura, & Rossi, 1992; Resnick & Jenkins, 2000). In the context of measuring SEEB in older adults, IRT also indicates where that information is provided on each of these two scales. In other words, is one scale better for measuring low levels of self-efficacy and the other higher levels when measuring self-efficacy in sedentary older adults? We hypothesized that the two different scales would be differentially reliable, the Resnick SEEB being more reliable than the Marcus SEEB, for measuring different levels of SEEB in these community-based older adult populations with arthritis.

It is also possible to use this information from the IRT analysis to gain a better understanding of how reliably change can be detected, and this is a unique contribution of this study. We looked at change in the Resnick SEEB; this scale was specially tailored to measuring self-efficacy in older adults.
Materials and Methods

Data from two large community-based randomized controlled trials of physical activity programs in adults with arthritis were analyzed. The first is the People With Arthritis Can Exercise (PACE) program with 346 participants from 18 sites across North Carolina. Functional capacity, symptoms, and psychological status were assessed at baseline and 8-week follow-up assessments (Callahan et al., 2008). The second, Active Living Every Day (ALED), a behavior-based program, was evaluated in 354 participants from 17 sites across the state (Callahan et al., 2007). For ALED, assessments were done at baseline and 20 weeks. The two SEEB scales collected at baseline and follow-up are the Resnick SEEB from ALED and the Marcus SEEB from PACE (Marcus & Owen, 1992; Marcus et al., 1992; Resnick & Jenkins, 2000). Approval for this secondary data analysis and the original PACE and ALED studies was granted by the University of North Carolina Biomedical Institutional Review Board. The risk involved in this study was determined to be no more than minimal, and additional consent was not required. The original PACE and ALED studies were conducted with the understanding of the consequences of taking part, and written consent was obtained from the participants.

The 5-item Marcus SEEB scale with 5-point Likert response levels assesses respondents’ confidence in their ability to be physically active despite common barriers (i.e., adverse weather, lack of time, etc.; Marcus & Owen, 1992; Marcus et al., 1992). A score is calculated by averaging responses to all five items, yielding a possible range of 1–5 points. Higher scores reflect higher levels of self-efficacy for physical activity. The Marcus SEEB has shown internal consistency across multiple CTT studies (Marcus & Owen, 1992; Marcus et al., 1992). Construct validity of the Marcus SEEB has been supported by repeatedly explaining a significant portion of variance in physical activity behavior again using CTT techniques (Kubiak, 2004; Oliver & Cronan, 2002).

The Resnick SEEB is a 9-item scale that measures self-efficacy to overcome barriers to physical activity (Resnick & Jenkins, 2000; Resnick & Spellbring, 2000). It has a 10-point response pattern with one representing not confident and 10 very confident. Scores from the Resnick SEEB are also created by averaging responses. A higher total score indicates higher self-efficacy for overcoming barriers to exercise. The Resnick SEEB has sufficient internal consistency, reliability, and validity in an older adult population as assessed by CTT (Resnick & Jenkins, 2000; Resnick & Spellbring, 2000).

Analyses

Both SEEB scales consist of multiple ordered responses. One widely used IRT model with this kind of data is the graded-response model (GRM; Samejima, 1969). The GRM parameter estimates were obtained from MULTILOG (Thissen, 1991). The version of the GRM used is unidimensional, assuming that one and only one construct is being measured by each SEEB scale. To evaluate the plausibility of this assumption, exploratory and confirmatory factor analyses were performed (EFA and CFA, respectively). The EFA analyses were conducted using CEFA, and the CFA analyses, using LISREL (Browne, Cudeck, Tateneni, & Mels, 2004;
Jöreskog & Sörbom, 2004). Both software packages use polychoric correlations, which correct for the categorical nature of the observed data (Wirth & Edwards, 2007). Ordinary least-squares estimation was used in the EFA, and diagonally weighted least-squares estimation was used in the CFA. The EFA results were evaluated based on the resulting eigenvalues (which are rough indicators of how much variance each factor accounts for) and factor loadings. The fit of the one-factor model was evaluated in the CFA framework using the root-mean-square error of approximation (RMSEA), comparative fit index (CFI), the goodness-of-fit index (GFI), and root-mean-square residuals (RMSR). Published rules of thumb (Hu & Bentler, 1999) suggest that RMSEA values less than .05 are good, .05–.08 are adequate, greater than .08 to .1 mediocre, and greater than .1 indicate unacceptable fit. Both CFI and GFI have lower cutoffs of .9 for reasonable fit, and RMSR values less than .1 are generally considered acceptable. From the IRT analysis we focus on the item parameters and standard-error curves to gain a better understanding of the performance of each scale. The Resnick SEEB was collected at baseline and a 20-week follow-up and was used to illustrate differences between summed scores and IRT scale scores.

Results

Factor Analysis

The number of participants providing complete data on the Marcus SEEB was 283, and for the Resnick SEEB was 312. Listwise deletion was used to eliminate missing data for the factor analyses. In both cases, the EFA strongly suggested the presence of one dominant factor. Both scree plots (which are simply plots of the eigenvalues) were indicative of one-factor solutions, and the second eigenvalues were appreciably smaller than the first (3.2 vs. 0.63 on the Marcus SEEB and 5.5 vs. 0.72 on the Resnick SEEB).

After the EFAs we conducted one-dimensional CFAs to gauge the extent to which a one-factor model fit the data for both scales. The Marcus SEEB one-dimensional model fit well (RMSEA = .03, GFI = 1.0, CFI = 1.0, RMSR = .03), and, although the Resnick SEEB did not fit quite as well, the overall levels of fit are still acceptable (RMSEA = .08, GFI = .99, CFI = .99, RMSR = .04). In both cases, the CFA results lend further support to the unidimensionality of these scales.

IRT Analysis

Item parameters for both scales are given in Table 1. The estimated parameters show that the item slopes are fairly variable, ranging from 1.73 to 3.89 for the Resnick SEEB and from 1.48 to 3.14 for the Marcus SEEB.

The standard-error curves for the Resnick and Marcus SEEBs are presented in Figure 1. The latent factor is along the x axis, and the standard error, on the y axis. The curves reflect the standard error that would accompany a score given at a particular level of exercise self-efficacy. For instance, scores between –1.6 and 1.2 on the Resnick SEEB are roughly equally precise, with a standard error near 0.22. By virtue of the way in which the scale is set in most applications of IRT, the latent construct is a standard normal distribution. This means that IRT scale scores
Table 1  Item-Response-Theory Parameter Estimates

<table>
<thead>
<tr>
<th>Resnick SEEB Item Parameters</th>
<th>$a$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
<th>$b_5$</th>
<th>$b_6$</th>
<th>$b_7$</th>
<th>$b_8$</th>
<th>$b_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The weather is bothering you.</td>
<td>1.73</td>
<td>-2.34</td>
<td>-1.83</td>
<td>-1.22</td>
<td>-0.91</td>
<td>-0.25</td>
<td>-0.03</td>
<td>0.22</td>
<td>0.72</td>
<td>1.25</td>
</tr>
<tr>
<td>2. You are bored by the program or activity.</td>
<td>2.51</td>
<td>-1.89</td>
<td>-1.52</td>
<td>-1.13</td>
<td>-0.83</td>
<td>-0.28</td>
<td>-0.02</td>
<td>0.21</td>
<td>0.73</td>
<td>1.20</td>
</tr>
<tr>
<td>3. You feel pain when exercising.</td>
<td>1.83</td>
<td>-2.20</td>
<td>-1.81</td>
<td>-1.28</td>
<td>-0.82</td>
<td>-0.17</td>
<td>0.13</td>
<td>0.52</td>
<td>1.18</td>
<td>1.72</td>
</tr>
<tr>
<td>4. You have to exercise alone.</td>
<td>2.19</td>
<td>-1.97</td>
<td>-1.50</td>
<td>-1.08</td>
<td>-0.77</td>
<td>-0.35</td>
<td>-0.10</td>
<td>0.17</td>
<td>0.72</td>
<td>1.13</td>
</tr>
<tr>
<td>5. You do not enjoy it.</td>
<td>2.93</td>
<td>-1.63</td>
<td>-1.37</td>
<td>-1.00</td>
<td>-0.73</td>
<td>-0.20</td>
<td>0.09</td>
<td>0.38</td>
<td>0.85</td>
<td>1.29</td>
</tr>
<tr>
<td>6. You are too busy with other activities.</td>
<td>2.19</td>
<td>-1.77</td>
<td>-1.50</td>
<td>-1.08</td>
<td>-0.76</td>
<td>-0.30</td>
<td>0.07</td>
<td>0.47</td>
<td>0.99</td>
<td>1.44</td>
</tr>
<tr>
<td>7. You feel tired.</td>
<td>3.10</td>
<td>-1.72</td>
<td>-1.43</td>
<td>-0.98</td>
<td>-0.66</td>
<td>-0.15</td>
<td>0.14</td>
<td>0.47</td>
<td>0.95</td>
<td>1.40</td>
</tr>
<tr>
<td>8. You feel stressed.</td>
<td>3.89</td>
<td>-1.57</td>
<td>-1.31</td>
<td>-0.99</td>
<td>-0.72</td>
<td>-0.18</td>
<td>0.01</td>
<td>0.32</td>
<td>0.74</td>
<td>1.15</td>
</tr>
<tr>
<td>9. You feel depressed.</td>
<td>3.06</td>
<td>-1.50</td>
<td>-1.18</td>
<td>-0.86</td>
<td>-0.62</td>
<td>-0.18</td>
<td>0.05</td>
<td>0.25</td>
<td>0.66</td>
<td>1.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marcus SEEB Item Parameters</th>
<th>$a$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
<th>$b_5$</th>
<th>$b_6$</th>
<th>$b_7$</th>
<th>$b_8$</th>
<th>$b_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. When I am tired.</td>
<td>2.15</td>
<td>-1.00</td>
<td>-0.11</td>
<td>1.16</td>
<td>2.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. When I am in a bad mood.</td>
<td>3.14</td>
<td>-1.07</td>
<td>-0.24</td>
<td>0.61</td>
<td>1.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. When I feel I don’t have time.</td>
<td>2.67</td>
<td>-0.98</td>
<td>-0.09</td>
<td>0.91</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. When I am on vacation.</td>
<td>1.48</td>
<td>-1.58</td>
<td>-0.64</td>
<td>0.41</td>
<td>2.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. When it is raining or snowing.</td>
<td>1.65</td>
<td>-0.91</td>
<td>-0.02</td>
<td>1.15</td>
<td>2.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. SEEB = self-efficacy for exercise behavior. Item parameters for full versions of both SEEB scales are shown. Slopes ($a$ parameters) describe the strength of the relationship between the item and latent construct. Higher slopes mean stronger relationships (analogous to a higher factor loading or higher correlation). The thresholds ($b$ values) are in a standard normal metric and describe the transition points between the available response options. At these points along the normal continuum the most likely response option changes.
are also in a standard normal metric. As the scores are in a standard normal metric, so too are the standard errors. The standard-error curve for the Marcus SEEB suggests that this scale provides reasonably precise scores between –1.2 and 2, but the scores become increasingly less precise beyond that level. For the Resnick SEEB, the standard-error curve indicates that the most precise scores from this scale are in the range of –2 to 1.5. As with the Marcus SEEB, beyond these points scores from the Resnick SEEB become less precise.

Marginal reliability is an IRT analog to CTT reliability coefficients such as alpha. These one-number summaries do not convey as much information as the standard-error curves but can be useful to get an overall sense of how reliable scores are. The marginal reliability was .86 for the Marcus SEEB and .95 for the Resnick SEEB.

**IRT and Change Scores**

Figure 2 plots changes in IRT scores (Time 1 – Time 2) for individuals (n = 83) with summed scores that changed (Time 1 – Time 2) less than 5 points. Summed scores on the Resnick SEEB range from 9 to 90, while IRT scores follow a standard
normal distribution (i.e., roughly 99% of scores will be between ±3). For summed-score changes ranging from −5 to +5, Figure 2 plots the corresponding change in IRT scores. For example, there are 12 people who had the same summed score at the follow-up as at the baseline assessment. Of those 12, 4 had identical response patterns, as well, which means their IRT scores are identical. The other 8 people had different response patterns at the follow-up than at the baseline. In two of these cases, that change in response pattern led to a substantial change in IRT-estimated scale score (.36 and .5 in the standard normal metric).

Discussion

As we hypothesized, the Resnick SEEB was slightly more precise for lower levels of self-efficacy, whereas the Marcus SEEB was more precise for higher levels of self-efficacy, in community-dwelling older adults. These findings can guide future researchers choosing a scale for a particular purpose. One example is that in an intervention focused on increasing physical activity in an assisted-living population (a potentially frail population) the Resnick SEEB is the better choice. A second
example is in the PACE and ALED studies; although the populations were primarily sedentary older adults, adults under 65 were not excluded, and the Marcus SEEB performed well in this population.

Whatever the context, the goal of measurement is to differentiate respondents according to their level on some construct of interest. The better able we are to achieve this differentiation, the more reliable our scores are likely to be. The IRT approach described here provides differentiation in three different ways. First, as demonstrated in Figure 2, the use of IRT scores increases the variability among scores. Second, the standard-error curve makes it easy to see that there is differentiation in the precision of the different scores. The standard-error curves summarize the precision with which a given scale can measure. Having standard errors in an easily interpretable metric is yet one more advantage of the IRT approach described here. The standard-error curves reflect the fact that in IRT models, precision is never assumed to be constant for all scores. Finally, by extending the ideas about precision to the study of change, an IRT-based approach allows us to begin to differentiate which scales might be useful for assessing what kinds of change. For example, neither scale examined here would be particularly good at detecting changes in self-efficacy in individuals who are more than 2 standard deviations above the mean. In regard to the IRT and change score, using the Resnick SEEB in a CTT approach would have missed substantial change.

To the extent that the points at a given summed-score difference vary along the y axis, the IRT scoring procedure is adding variability that the summed-score model ignores.

The Resnick SEEB has 10 response patterns, and there was initial concern that some of those categories might have to be collapsed to perform the IRT analysis. We were surprised to discover that despite the large number of response categories, respondents seemed to be using all of them with some regularity. The second lowest response category (2) was typically the least endorsed option, but even this category was chosen by approximately 4% of the respondents.

The current study is not without its limitations. The fit of the one-factor model for the Resnick SEEB was within published norms for acceptable fit but was not as good as we might have liked. Examination of the model revealed no obvious areas for improvement. It is difficult to know whether the Marcus and Resnick scales met the three features of self-efficacy theory (Forsyth & Carey, 1998). There was a relationship between these SEEB scales and physical activity interventions in the parent studies for those who completed the interventions. This relationship was not significant in the PACE study using the Marcus SEEB, but it was significant in the ALED using the Resnick SEEB, possibly supporting our hypothesis here that the Resnick SEEB is more precise at capturing lower levels of self-efficacy (Callahan et al., 2008; Callahan et al., 2007). Although the ALED and PACE populations were similar, they were not identical.

We identified the relative strengths of two SEEB scales using IRT. For both the Marcus SEEB and the Resnick SEEB, the IRT analyses indicated that all items on each scale were contributing useful information to the total score. These scales are precise and reliable measures over fairly similar ranges of self-efficacy and should be useful in identifying people with arthritis.
References


