HEALTH EDUCATION RESEARCH
Theory & Practice

Vol.21 (Supplement 1) 2006
Pages i110–i120
Advance Access publication 22 August 2006

Introducing equating methodologies to compare test scores from two different self-regulation scales

Louise C. Mâsse[1]*, Diane Allen[2], Mark Wilson[2] and Geoffrey Williams[3]

Abstract

Standardizing the measurement tools that researchers use to assess the effectiveness of interventions would strengthen our ability to compare results across studies. In practice, however, standardization is difficult to implement, in part, because researchers prefer to use measurement tools that focus specifically on the components of their interventions. This paper demonstrates the usefulness of item response modeling linking methodology in comparing groups of participants who were administered different scales intended to measure the same underlying constructs. The Treatment Self-Regulation Questionnaire (TSRQ) as it relates to diet improvement provided the empirical application to demonstrate how two different scales that measure the same construct can be compared. The results showed that two eight-item TSRQ scales can be linked if they have at least four items in common. As expected, varying the number of linking items did not affect the reliability of the results; however, it significantly affected the relative rating with respect to the 15-item scale.

In health behavior and health education research, linking methodologies can be used to compare results across studies that use slightly different versions of a scale to measure the same construct.

[1]Department of Pediatrics, University of British Columbia, Centre for Community Child Health Research, L408, 4480 Oak Street, Vancouver, BC V6H 3V4, Canada, [2]Graduate School of Education, University of California Berkeley, Berkeley, CA 94720, USA and [3]Clinical and Social Psychology Department, University of Rochester, Rochester, NY 14627, USA
***Correspondence to: L. C. Mâsse.**
E-mail: lmasse@cw.bc.ca

Introduction

Standardizing the measurement tools that researchers use in the health domain has several benefits, including increasing our ability to compare results across studies and integrate results from various studies into a meta-analysis. Standardizing measurement tools is difficult to implement, however, because researchers prefer to use tools that focus on the specific components of their studies or interventions. In practice, it is not uncommon for researchers to modify already validated scales for their particular research context. When study participants use different scales, however, even if the scales purportedly measure the same construct, the raw scores on the different scales will not be comparable if the scales are not based on the same metric. Even if the scales use the same response format and contain the same number of items, the resulting scores can differ because the scales may emphasize a different aspect of the construct. Standardizing methodologies can help to address some of these issues by ensuring that different scales use the same metric and thus are comparable. Although there are several test-equating applications in health outcomes research, fewer applications of these methodologies exist in health behavior and health education research.

Test equating is the process used to establish a mathematical relationship between two scales so

doi:10.1093/her/cyl088

that the scores on these scales are based on the same metric and thus are comparable [1]. In lay terms, equating consists of establishing equivalence scores on two different scales that measure the same construct. The most stringent form of equating has five assumptions: (i) the scales measure the same construct; (ii) the scales have similar measurement errors; (iii) the distribution of scores remains the same across scales for a participant of a given level (i.e. equal distribution between the two scales); (iv) equating from Scale 1 to Scale 2 is equivalent to equating from Scale 2 to Scale 1 (i.e. changing the order in which equating is done leads to the same results) and (v) the equating relationship is group invariant (i.e. the equating relationship does not change across groups) [2]. As not all these assumptions are likely to be tenable in health research, we use a less stringent form of equating (consisting of relaxing assumptions (i) and (ii)), which is referred to as 'linking' or 'scaling' [3]. Linking allows two instruments that measure the same construct to use the same metric, even when their precision and emphasized level of construct differs.

Three designs can be employed to collect data for linking purposes: (i) a random-group design, in which Scales 1 and 2 are randomly assigned to participants; (ii) a single-group design with counterbalancing, in which Scales 1 and 2 both are assigned to the participant but the order of administration is counterbalanced and randomly assigned and (iii) the common item nonequivalent groups design, in which a subset of common items (internal or external to the scale) is administered to both groups of participants, but each group receives a subset of items that the other group does not receive [1]. Linking can be achieved under classical test theory (CTT) or item response modeling (IRM). Selecting the appropriate procedure (CTT versus IRM procedures) depends on the tenability of the assumptions associated with the procedure and design employed to collect the data. The common item nonequivalent groups design (design (iii)) is the least burdensome to participants and is the most likely design to be used in health behavior and health research. When this design is employed, CTT is not well suited to handle such data given that the scales are not randomly assigned to the groups. However, it is particularly straightforward under IRM to link scales as the item parameters (when the model assumptions are verified) are assumed to be invariant across groups. Item parameters describe the discrimination and difficulty or location of the item on the latent construct. This paper will demonstrate how IRM can be used to link scores of participants in groups that have been administered different scales with a subset of common items. Specifically, the purpose of this paper is to (i) demonstrate how linking methodology can be used to link test scores from two different scales that include a common set of items and (ii) discuss design considerations needed to achieve valid test equating, with a specific focus on understanding the impact of varying the number of linking items. The Treatment Self-Regulation Questionnaire (TSRQ), developed by Ryan and Connell [4] for education and adapted by Williams *et al*. [5] for health behavior, is used to illustrate the usefulness of linking methodologies.

Methods

Participants

Data collected as part of the Behavior Change Consortium (BCC) were analyzed for this paper [6]. Two of the 15 BCC sites provided data for the analyses: the Oregon Health Sciences University (OHSU) and the University of Rochester (UR). The OHSU enrolled firefighters in a study aimed at improving dietary and physical activity behaviors. The UR study enrolled adult smokers in a tobacco-dependence treatment and diet intervention study [7].

The study design, recruitment and detailed description of the intervention for the UR study have been reported elsewhere [8]. In brief, people who smoked five or more cigarettes per day, were 18 years of age or older, read and spoke English, had no history of a psychotic illness (depression and anxiety were allowed), had a life expectancy of 18 months and planned to live in the greater Rochester, NY, area for 18 months were recruited through newspaper ads and signs in physician offices to participate in a study about 'smokers' health'.

The OHSU study targeted firefighters stationed at five fire departments in Oregon and Washington to participate in a study about promoting healthy lifestyles for firefighters [9]. Firefighters were recruited through personal contact and informational video distributed at the fire station. In total, 696 firefighters met inclusion criteria for the study. Baseline data were collected from firefighters while on duty prior to randomization for the study.

Baseline data for both sites were analyzed for this paper. As the firefighters were predominantly male, only the data from males in both studies were analyzed so that any effects of gender would not confound the group results. The OHSU sites had 627 males who completed the TSRQ at baseline; 355 of the UR males had available data. The demographic characteristics of the participants are presented in Table I.

Table I. *Demographic characteristics of participants by groups (n = 982)*

	OHSU (n = 627)	UR (n = 355)
Race/ethnicity (%)		
White, not Hispanic	90.1	81.1
Black, not Hispanic	2.1	13.0
Hispanic	3.1	2.2
Other	4.7	3.7
Age (mean, SD)	40.1 (8.9)	46.1 (12.0)
Marital status (%)		
Married	79.8	45.1
Not married	10.2	26.8
Living with partner	2.9	9.6
Divorced/separated/widowed	7.1	18.6
Education (%)		
Less than high school	0.0	7.4
High school diploma	11.6	31.3
Some college	16.5	37.7
College graduate	21.6	23.6
Income (%)		
<$40 000	4.0	48.1
$40 000–79 000	55.3	38.6
⩾$80 000	40.7	13.3

Treatment self-regulation questionnaire

Self-determination theory (SDT) details the motivational basis for self-regulation of human behavior and focuses on the concept of autonomy [10]. SDT distinguishes between autonomous, controlled and amotivated reasons for behavior. The TSRQ asks participants to endorse different reasons for behavior and thus records the extent to which regulatory processes are self-determined. Autonomous motivation for dietary change indicates that people experience a sense of choice and a sense of volition about following a specific diet. Controlled regulation, in contrast, indicates that people feel pressured or coerced by themselves (intrapersonal control) or others (interpersonal control) to follow a specific diet. Amotivated regulation occurs when people feel that there is no connection between their efforts to follow a specific diet and their health outcomes. Internalization is the process through which motivation becomes more autonomous. An increase in TSRQ autonomy over time reflects internalization and is expected to result in sustained healthy dietary behavior. Autonomous and controlled motivations for changing diet were assessed using six items each; amotivation was assessed by three items (see Table II). Participants responded to each item on a scale of 1 (strongly disagree) to 7 (strongly agree). An example of an autonomous motivation for eating a healthy diet is 'I feel that I want to take responsibility for my own health'. An example of a controlled motivation is 'I would feel guilty or ashamed of myself if I did not eat a healthy diet'. An example of an amotivated reason is 'I really don't think about it'.

Each subscale score is generated by computing a mean for the items that relate to that subscale. Reliability was assessed with Cronbach's alpha. For the autonomous subscales, Cronbach's alpha was 0.91 in the Rochester and OHSU samples, for controlled motivation it was 0.83 and 0.81 for the Rochester and the OHSU sample, respectively, and for amotivation it was 0.55 and 0.57, respectively.

Protocol

Baseline assessments

At UR, all participants completed at baseline, questionnaires that assessed demographic information, medical history, smoking history and intention to

Table II. *Items administered to the reference group (OHSU) and comparison group (UR) by linking conditions*[a]

Items[b]	Reference group (OHSU)	Comparison group conditions[c] (UR)					
		8_8	8_1	8_2	8_3	8_4	8_5
1. I feel that I want to take responsibility for my own health.	X	**X**			**X**	**X**	**X**
2. I would feel guilty or ashamed of myself if I did not eat a healthy diet.			X				
3. I personally believe it is the best thing for my health.	X	**X**	**X**	**X**	**X**	**X**	**X**
4. Others would be upset with me if I did not.	X	**X**				**X**	**X**
5. I really don't think about it.			X	X	X	X	X
6. I have carefully thought about it and believe it is very important for many aspects of my life.			X	X	X	X	X
7. I would feel bad about myself if I did not eat a healthy diet.	X	**X**		**X**	**X**	**X**	**X**
8. It is an important choice I really want to make.	X	**X**					**X**
9. I feel pressure from others to do so.			X	X	X		
10. It is easier to do what I am told than think about it.			X	X			
11. It is consistent with my life goals.	X	**X**					
12. I want others to approve of me.	X	**X**					
13. It is very important for being as healthy as possible.			X	X	X	X	
14. I want others to see I can do it.			X	X	X	X	X
15. I really don't know why.	X	**X**					

[a]X = items administered, and **X** (bolded) = items that were administered to both groups.
[b]Items 2, 4, 7, 9, 12 and 14 measure autonomous motivation. Items 1, 3, 6, 8, 11 and 13 measure controlled motivation. Items 5, 10 and 15 measure amotivation.
[c]Conditions 8_8 = eight items with eight common items, 8_1 = eight items with one common item; 8_2 = eight items with two common items, 8_3 = eight items with three common items; 8_4 = eight items with four common items and 8_5 = eight items with five common items.

quit smoking in the next 30 days. Participants also had their blood pressure measured. In addition, participants completed the Fagerstrom Addiction Severity Scale [11] and the TSRQ. At OHSU, participants completed self-report questionnaires at the same time as a battery of objective physical measures (e.g. body mass index, blood pressure, glucose and lipid profiles) were obtained.

Simulated conditions

Both groups received all 15 TSRQ items. We simulated conditions in which different groups receive different sets of items by selecting the items for which the responses were analyzed as a set and 'eliminating' all other item responses for a particular analysis. For example, we simulated a condition in which both groups received eight items with only four items in common to demonstrate how linking is performed and results are interpreted. The items that were assigned to OHSU and UR in this simulation are described in Table II (see the condition with four linking items, 8_4). To evaluate the impact of varying the number of linking items on the accuracy of the results, we simulated five additional conditions. Overall, the simulation assigned eight items per group, with one, two, three, four, five and eight common items per condition; referred herein as conditions 8_1, 8_2, 8_3, 8_4, 8_5 and 8_8, respectively. Table II shows which items were assigned to the two groups under the various conditions. Finally, to evaluate the impact of choosing any one particular item as a single linking item, we simulated seven conditions in which the single linking item used in condition 8_1 differed (note: Table II shows only one of those seven conditions). Each participant in a group was simulated to receive the same set of items as the rest of the group. The common item design was simulated in all conditions, meaning that the linking items were included in the scale score. In all simulated conditions, the items were chosen to balance as much as possible the content of the items

included in each form and to represent the areas addressed by the three subscales.

Analyses

The partial credit model [12] was used for all the analyses, and all IRM analyses were conducted using ConQuest software [13]. Linking can be achieved in two ways with the common items design. Either a separate or concurrent calibration can be performed to link items on two different scales that measure the same construct and have a subset of common items [14]. A separate calibration consists of computing the item parameters for the two groups separately, even for the common items. A linear transformation then is used to link the metrics of the groups by using the common items as an anchor for the linking. Alternatively, a concurrent calibration can be performed by simultaneously estimating the item parameters of both groups in the same computer run. Both groups are combined, and the items that were not administered to a given group are considered as missing at random; item parameters for both groups are estimated simultaneously and thus are based on a comparable metric. Given that a concurrent calibration is considered more precise than a separate calibration [1], we used the concurrent calibration method in this paper.

For the linking application, we compared the item parameters estimated for the full scale (15 items administered to both groups, condition 15_15) with those estimated for the reduced scale (conditions 8_1 to 8_8). To compare the impact of varying the number of items used for linking the scales on the construct estimate, we compared the reliability of the construct with the full scale and compared its impact on the scale estimate for each individual (e.g. self-regulation estimate). The proportion of aberrant estimates was computed by comparing individual estimates obtained from the reduced scale with those obtained from the full scale. If the estimate based on the six linking conditions was more than ± 1 standard error (SE) away from the estimate obtained from the full scale, it was considered to be meaningfully different and an aberrant case. The relative increase in aberrant cases was calculated to represent the percentage of aberrant cases resulting from nonequivalent scales. For example, the relative increase of aberrant cases for condition 8_4 equaled the difference between the percent aberrant cases for 8_8 minus the percent aberrant cases for condition 8_4, thus eliminating the portion of aberrant cases that resulted from shortening the scale. The expected *a posteriori* reliability for the individual estimate (e.g. person separation reliability [15]) was computed in logits (the log of the odds of a particular set of responses) for each condition, along with the correlation between the linking conditions and the full-scale estimate.

A unidimensional Rasch model was fitted to the data. Unidimensionality is verified when only one main dimension exists, which does not eliminate the presence of minor dimensions including subscales [16]. The TSRQ includes a number of correlated subscales which are assumed to measure one main dimension of self-determination. To verify the unidimensional assumption, a principal components analysis was conducted using SPSS. Reckase's [16] criterion was used to confirm dimensionality: if the first dimension explained 20% of the total variance, the unidimensionality assumption was assumed to hold. Ultimately, unidimensionality was verified by evaluating the fit of the Rasch model by examining the weighted mean square indices for the item parameters. Item parameters with weighted mean square indices of <0.75 or >1.33 and t-statistics of greater than ± 1.96 indicated a poor fit [17].

Common items used for linking are assumed to function similarly in the two groups, meaning that the probability of selecting a given response option is the same for individuals who have the same scale score (i.e. the item parameters are stable across groups). This type of analysis is referred to as differential item functioning (DIF), and items that exhibit severe DIF cannot be used for linking [18]. The DIF analyses were conducted in ConQuest on the 15-item scale. An alpha of 0.05 served to determine if overall DIF existed. To evaluate which items exhibited DIF, the ratio of the item parameter location and SE served to locate the significant

differences, a ratio of greater than ±1.96 was deemed significant. Using the Educational Testing Service standards for effect size (as modified for Rasch-scaled instruments), 'negligible', 'moderate' and 'large' DIF are defined as a difference in the item parameter location of <0.426, 0.426–0.638 and >0.638, respectively [19]. Items that exhibit moderate to large amounts of DIF are considered inappropriate for linking purposes, but a negligible amount of DIF has not been found to significantly impact linking. In IRM, the items and scores are on the same metric; therefore, examination of the item parameter location allows one to assess where the item is providing most information on the scale score.

Results

Linking application

The unidimensionality assumption was verified for condition 8_4, and the linking application was fully demonstrated for this condition. The principal component analysis results indicated that the eight items assigned to OHSU and UR condition 8_4 explained 40.3 and 41.6% of the total variance, respectively. These results suggested that the eight-item test was sufficiently unidimensional to proceed with a unidimensional IRM linking application. As we compared the item parameters obtained via linking with the reference condition (in which all items were taken by both groups), we also verified the unidimensionality of the 15-item scale. The principal components analysis results showed that the first dimension explained 33.0% of the total variance. Again, this result demonstrated that the scale was sufficiently unidimensional to proceed with unidimensional IRM, although it did not preclude the scale from having minor dimensions. Examining the IRM fit indices (weighted mean square indices and *t*-statistics) provided another way to assess the unidimensionality of the scale. For all conditions, Items 5 and 15, both amotivated items, were misfitting according to their high weighted mean square indices and significant *t* values, which indicated that responses on these items did not match the patterns of responses on the other items. These items were retained, however, as they represented an important option for participants in a balanced set of items chosen for the simulated conditions.

The DIF analysis on the 15-item scale was significant and indicated that DIF indeed was present (χ^2 = 56.073, df = 14, P = 0.000). An examination of the item parameter location estimate and SE ratio for each item revealed that seven items had significant DIF (results not shown). The magnitude of DIF, however, was negligible for all items (the difference in item parameter locations ranged from 0.16 to 0.44), which suggested that any item on the scale could be used to anchor the linking. To simulate equivalent scales with appropriate content coverage, one of the items with negligible DIF was used in the linking application (Item 7). Because this item had the lowest difference in item parameter location (0.16) between groups, it was not expected to impact the linking because negligible DIF has been found to have little impact on linking. None of the other items with negligible DIF was used in any of the linking conditions.

Table III presents the linked item location parameter estimates and SEs for condition 8_4 (eight-item scale with four common items) and condition 15_15 (15 overlapping items). The item location parameters for condition 15_15 ranged from −0.85 to 0.82. As the name implies, item location parameters indicate where items are located on the self-determination continuum. The ConQuest software anchors the continuum at 0, which means that the mean of the item location parameters is equal to 0 (this is a common assumption across all the analyses). Positive item location parameters indicate that self-determination items are more difficult to endorse, and vice versa. The six items with negative item location parameters (easier items to endorse) were Items 1, 3, 6, 8, 11 and 13, which were used in the self-determination application to assess the amount of autonomous motivation. Given their location on the continuum, these items are assumed to be easier to endorse than the other items for those with high levels of self-determination. This is consistent with previous

Table III. *Item location parameter estimates and SEs for various conditions*

Items	Item location									
	Parameter estimate by conditions					SE by conditions				
	15_15	8_8	8_4			15_15	8_8	8_4		
			Eight-item scale with four common items					Eight-item scale with four common items		
	15 overlapping items	Eight overlapping items	OHSU items	Common items	UR items	15 overlapping items	Eight overlapping items	OHSU items	Common items	UR items
1	−0.85	−0.88		−0.85		0.028	0.028		0.028	
2	0.34	NA[a]	NA		NA	0.020	NA	NA		NA
3	−0.87	−0.91		−0.87		0.029	0.029		0.029	
4	0.69	0.71		0.69		0.024	0.024		0.024	
5	0.34	NA	NA		0.26	0.020	NA	NA		0.029
6	−0.46	NA	NA		−0.43	0.022	NA	NA		0.032
7	0.12	0.12		0.12		0.019	0.020		0.019	
8	−0.57	−0.60	−0.69		NA	0.024	0.025	0.034		NA
9	0.77	NA	NA		NA	0.024	NA	NA		NA
10	0.82	NA	NA		NA	0.026	NA	NA		NA
11	−0.43	−0.45	−0.57		NA	0.023	0.023	0.032		NA
12	0.73	0.76	0.76		NA	0.023	0.023	0.029		NA
13	−0.83	NA	NA		−0.86	0.028	NA	NA		0.045
14	0.43	NA	NA		0.35	0.020	NA	NA		0.030
15	0.64	0.67	0.91		NA	0.023	0.023	0.033		NA

[a]NA = Response to this item was not analyzed in this simulated condition.

experience with this scale as autonomy items typically have a higher mean than the other subscales items. Linking places all the items (common and group-specific items) on the same underlying continuum distribution, which allows the relative location of the items to be comparable between groups. In condition 8_4, items that were administered to both groups were expected to have similar (or identical) item location parameters as the 15_15 condition, and other items were expected to deviate slightly. As shown in Table III, the common items (Items 1, 3, 4 and 7) in condition 8_4 had identical item location parameters to the 15_15 condition. The difference in item location parameter was largest for Item 15, an item from the amotivated subscale, where the difference between item location parameter was 0.27. Among the group-specific items, the differences in item location parameters for five of the eight items were significant (difference was >2 SE in absolute value from the 15_15 parameter estimates). However, the differences in item location parameter can be classified as negligible, according to the Educational Testing Service standards for effect size [19], as these differences are all <0.426. As reducing the number of items administered to each group is expected to impact the item location parameters, we assessed this impact by administering the same eight items to both groups (condition 8_8). All the items that were administered to the OHSU group in condition 8_4 were administered to both groups. The results (Table III) revealed that the item location parameter estimates were more similar between conditions 8_8 and 15_15 than they were between conditions 8_4 and 15_15.

The impact of slight variations in item location parameter estimates on the group estimates is presented in Table IV. Table IV shows the mean and standard deviation (SD) estimates for the self-determination scale for conditions 15_15, 8_8 and 8_4 by group. Results show that linking the groups is important in preserving observed group differences. The unlinked estimate for condition 8_4 slightly affected the mean and SD, whereas the groups' mean appeared to be similar in the unlinked condition. Note that group differences under the linked conditions with eight items were equal to 0.08 logits for both the 8_8 and 8_4 conditions (a logit is computed by taking the natural logarithm of the odds of obtaining this set of raw scores; in these Rasch analyses, the scale continuum had a mean of 0 and unconstrained SD). On a 48-point item scale, 0.08 logits correspond to an approximately two-point difference between the groups. In this case, this difference is not likely to be statistically and meaningfully different; however, group differences in other testing situations could be larger.

Table IV. *Mean and SD estimates (in Logit[a]) for the self-determination scale by conditions and groups (OHSU and UR) for linked and unlinked data*

	Conditions					
	15_15[b]		8_8[b]		8_4[b]	
	15 overlapping items		Eight overlapping items		Eight-item scale with four common items	
	OHSU	UR	OHSU	UR	OHSU	UR
Linked group estimate: Mean (SD)	−0.02 (0.34)	−0.13 (0.48)	0.10 (0.41)	0.02 (0.39)	0.13 (0.45)	0.05 (0.41)
Unlinked group estimate: Mean (SD)	—	—	—	—	0.14 (0.52)	0.13 (0.36)

[a]A logit is computed by taking the natural logarithm of the odds of obtaining this set of raw scores; in these Rasch analyses the scale continuum has a mean of 0 and unconstrained SD.
[b]A 0.04 Logit difference is equivalent to approximately three points in the 15-item scale (90 points possible).
[c]A 0.10 Logit difference is equivalent to approximately two points in the eight-item scale (48 points possible).

In addition to evaluating the impact on the group estimate, we evaluated the impact of linking on the case estimates by verifying the relative ranking of the participants using condition 15_15 as the referent group. Case estimates obtained under conditions 8_8 and 8_4 were compared with those obtained under condition 15_15. Case estimates that differed from condition 15_15 by more than ± 1 SE were considered aberrant; an aberrant case is defined as one in which a participant's level of self-determination differs meaningfully from condition 15_15. Reducing the scale from 15 to eight items, with eight linking items, resulted in 20.1% aberrant cases. In contrast, the impact of reducing the number of items from 15 to eight items with only four linking items resulted in 23.9% aberrant cases. Therefore, the relative increase in the percentage of aberrant cases as a result of administering a total of four different items to the two groups is ~3.8% (23.9 − 20.1%). The percentage of aberrant cases increased to 33.5% in condition 8_1. When investigating whether the choice of a particular single linking item makes a difference, aberrant item percentages ranged from 26.2 to 36.0%. Finally, as expected, scale reliability was reduced when there were fewer items in the scale—0.81 for 15 overlapping items and 0.64 for eight overlapping items. The reliability of conditions 8_8 and 8_4 was the same (0.64). Overall, the results showed that the group and case estimates, as well as scale reliability, were similar between conditions 8_8 and 8_4, which suggests that linking can be useful when comparing groups that have been administered different items.

Linking designs

The impact of varying the number of linking items on the percentage of aberrant cases is shown in Figure 1. Condition 15_15 served as the referent group for all the comparisons. As indicated above, the percentage of aberrant cases for condition 8_8 provided the impact of reducing the scale from 15 to eight items. Conditions 8_4 and 8_5 were the only two conditions in which the relative percentage of aberrant cases differed from the 8_8 condition by <5%. This suggests that the reliability

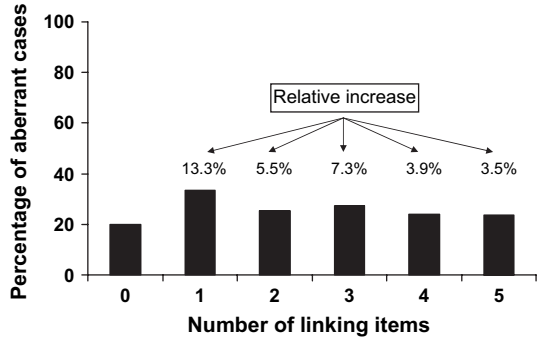

Fig. 1. Percentage of aberrant cases for the eight-item scale with 0, 1, 2, 3, 4 and 5 linking items and relative increase using the 0-item linking condition as the reference condition.

of the estimate may be compromised by using fewer than four linking items with an eight-item scale. Note that having only one linking item increased the percentage of relative aberrant cases by 13.3%. Post hoc analyses were conducted to determine if varying the item that served as the link could explain the sharp increase in the percentage of aberrant cases. Figure 2 presents the percentage of relative aberrant cases when the single linking item varied. The percentage of relative aberrant cases increased when Item 1 served as the linking item; it decreased when Item 7 was used for linking. Omitting Item 7, the relative increase (compared with condition 8_8) in the percentage of aberrant cases varied from 9.6 to 15.9% when one item served as the link. The impact of varying the number of items on the reliability index was minimal; reliability was found to range from 0.62 to 0.65 (Figure 3).

Discussion

This paper demonstrates how IRM linking methodology can be used to compare shorter versions of the TSRQ and the impact that varying the number of items used in the linking process can have on data accuracy. We showed that two simulated eight-item TSRQ scales can be linked reasonably well if the scale versions have at least four items in common. Four linking items were deemed necessary to

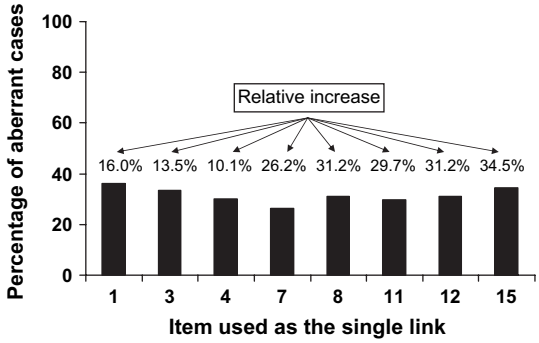

Fig. 2. Percentage of aberrant cases for the eight-item scale varying the item used as the linking item and relative increase using the 0-item linking condition as the reference condition.

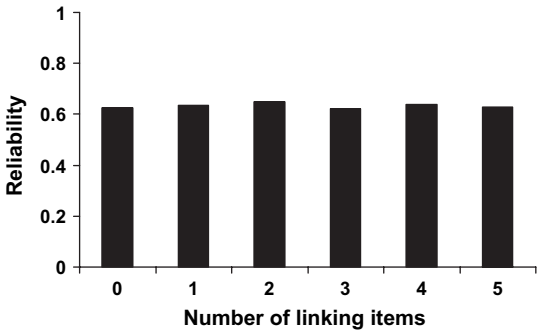

Fig. 3. Reliability index for the eight-item scale with 0, 1, 2, 3, 4 and 5 linking items.

keep the relative increase in percentage of aberrant cases from going >5%; meaning to maintain an adequate relative rating with respect to the 15-item scale. Because the two versions of the TSRQ were created so as to balance the number of items that assessed autonomous and controlled motivation and amotivation, it is possible that other versions could have required more or fewer linking items. Varying the number of linking items did not affect the reliability of the scale; this was somewhat expected because test reliability is affected more by the number of items included in the scale [20]. When comparing two groups that have taken two versions measuring the same construct, the scales must be linked to accurately assess group differences. In our application, linking had a minimal impact on the group estimate; however, this finding is not necessarily generalizable. For example, with longer scales and when the content is not as well counterbalanced as in this application, linking could have a greater impact on the group estimate. Overall, this paper demonstrates that linking can be useful when comparing groups that have taken different versions of the TSRQ.

It is important to consider some methodological issues that relate to linking. The purpose of linking as described in this paper is to enable the comparison of participants who have taken different versions of a scale. Linking is expected to decrease measurement errors in such comparisons. Violating the underlying assumptions of the linking process, however, may increase measurement errors; it may be best in such situations not to link the scales. In addition, the linking process can affect both random and systematic errors. Random errors are associated with sampling; using large samples is one way to control random errors [1, 21]. Controlling for systematic errors is more complex. Verifying assumptions that underlie the model used in the linking process is one way to minimize these errors. Prior to selecting a statistical model for the linking process, it also is important to consider the design that will be used to collect data, determine how many items will be used to link the scales, which items are most appropriate to serve as links between two scales, which group of participants is more suited for use in the linking process and sample size needed with a given statistical model. Clear decision-making rules for these issues do not exist, however. As there are no clear criteria to guide such analyses, it may be advantageous for researchers to conduct a pilot study to identify the best linking design for a given situation and to validate the linking process.

Linking methodologies can be of value to health behavior and health education research, as it is quite common for researchers to use slightly different scales to measure the same construct. Linking results would improve our ability to compare results across studies. It may be useful to establish normative scales for constructs of interest. Establishing normative scales would provide a standard

for the linking process and ensure that all studies use the same normative linking items. Identifying normative items for a given construct will require effort; once these items are identified, however, it will improve our ability to compare different studies by allowing study-specific items to be included in a given scale.

An alternative to identifying normative linking items is to develop an item bank of calibrated items. The methodology and procedures for using repositories of calibrated items to link different versions of a construct are well developed but have received little attention in health behavior and health education research. Item banking would provide increased flexibility for linking purposes. At present, the major obstacle to using item-banking methodologies relates to the process of developing and managing item banks [14]. The methodology to equate scales that measure the same construct is available, and health behavior and health education researchers would benefit from using these methods when comparing studies.

Acknowledgements

This project was conducted while L.C.M. was at the National Cancer Institute. Support for this project was provided by the National Cancer Institute contract number 263-MQ-319958. The authors would like to thank the BCC for providing the data that were used in the analyses. The views presented in this paper represent those of the authors and not those of the National Cancer Institute.

Conflict of interest statement

None declared.

References

1. Kolen MJ, Brennan RL. *Test Equating: Methods and Practices*. New York: Springer-Verlag, 1995.
2. Bjorner JB, Kosinski M, Ware JE. Using item response theory to calibrate the headache impact test (HIT™) to the metric of traditional headache scales. *Qual Life Res* 2003; **12**: 981–1002.
3. Dorans NJ. Scaling and equating. In: Wainer H, Dorans NJ, Eignor D, Flaugher R, Green BF, Mislevy RJ, Steinberg L, Thissen D (eds). *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, 135–58.
4. Ryan RM, Connell JP. Perceived locus of causality and internalization: examining reasons for acting in two domains. *J Pers Soc Psychol* 1989; **57**: 749–61.
5. Williams GC, Grow VM, Freedman ZR *et al*. Motivational predictors of weight loss and weight-loss maintenance. *J Pers Soc Psychol* 1996; **70**: 115–26.
6. Ory MG, Jordan PJ, Bazarre T. The behavior change consortium: setting the stage for a new century of health behavior-change research. *Health Educ Res* 2002; **17**: 500–11.
7. Williams GC, McGregor HA, Sharp D *et al*. Testing a self-determination theory intervention for motivating tobacco cessation: supporting autonomy and competence in a clinical trial. *Health Psychol* 2006; **25**: 91–101.
8. Williams GC, Minicucci DS, Kouides RW *et al*. Self-determination, smoking, diet and health. *Health Educ Res* 2002; **17**: 512–21.
9. Moe EL, Elliot DL, Goldberg L *et al*. Promoting healthy lifestyles: alternative models' effects (PHLAME). *Health Educ Res* 2002; **17**: 586–96.
10. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol* 2000; **55**: 68–78.
11. Fagerstrom K-O, Schneider NG. Measuring nicotine dependence: a review of the Fagerstrom Tolerance Questionnaire. *J Behav Med* 1989; **12**: 159–82.
12. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 198; **49**: 359–81.
13. Wu ML, Adams RJ, Wilson MR. *ACER ConQuest: Generalized Item Response Modeling Software*. Melbourne: The Australian Council for Educational Research, Ltd, 1988.
14. McHorney CA, Cohen AS. Equating health status measures with item response theory. *Med Care* 2000; **9**: S43–59.
15. Wright B, Masters G. The measurement of knowledge and attitude. *Research Memorandum No. 30*. Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago, 1981.
16. Reckase M. Unifactor latent trait models applied to multifactor tests: results and implications. *J Educ Stat* 1979; **4**: 207–30.
17. Adams RJ, Khoo ST. *Quest*. Melbourne: The Australian Council for Educational Research, Ltd, 1991.
18. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
19. Paek I. Investigation of differential item function: comparisons among approaches, and extension to a multidimensional context. *Unpublished PhD Dissertation*. Berkeley, CA: University of California, 2002.
20. Nunnally JC, Bernstein IH. *Psychometric Theory*, 3rd edn. New York: McGraw-Hill, 1994.
21. Masters GN, Keeves JP. *Advances in Measurement in Educational Research and Assessment*. New York: Elsevier, 1999.

Received on February 16, 2006; accepted on July 19, 2006