HEALTH EDUCATION RESEARCH
Theory & Practice

Vol.21 (Supplement 1) 2006
Pages i98–i109
Advance Access publication 23 October 2006

Does participation in an intervention affect responses on self-report questionnaires?

Tom Baranowski[1]*, Diane D. Allen[2], Louise C. Mâsse[3] and Mark Wilson[2]

Abstract

There has been some concern that participation in an intervention and exposure to a measurement instrument can change participants' interpretation of the items on a self-report questionnaire thereby distorting subsequent responses and biasing results. Differential item functioning (DIF) analysis using item response modeling can ascertain possible differences in item interpretation by testing for differences in item location between groups. The DIF for treatment versus control group differences at post-intervention assessment and the Time 1 and Time 2 differences in a control group were analyzed using data from a dietary change intervention trial for Boy Scouts. The measures included fruit and vegetable (FV) frequency of consumption, preferences and self-efficacy. Treatment–control group DIF at post-intervention assessment was detected in a higher percentage of items for FV frequency than for preference or self-efficacy. Time 1 to Time 2 differences in items for the control group were detected in one item for each of the three scales. Further research will need to clarify whether the obtained DIFs reflected true changes in frequency, preference or self-efficacy or some reinterpretation of items by participants following an intervention or merely after previous exposure to the measure.

[1]Department of Pediatrics, Baylor College of Medicine, Houston, Texas 77030, USA, [2]College of Education, University of California Berkeley, Berkeley, California 94720, USA and [3]University of British Columbia, Department of Pediatrics, Centre for Community Child Health Research, L408-4480 Oak Street, Vancouver, BC V6H 3V4, Canada
*Correspondence to: T. Baranowski.
E-mail: tbaranow@bcm.tmc.edu

Introduction

Theory-based behavioral interventions attempt to change selected behaviors through mediating variables [1–3]. Common psychosocial mediating variables in behavior change programs include preference [4] and self-efficacy [5]. Participants commonly reveal both behavioral outcomes and psychosocial mediators through self-report methods in intervention evaluations.

Some investigators have been concerned that encountering questions in self-report measures changes the participants' understanding of the behaviors targeted by the intervention and the factors influencing participants to do (or not to do) the behaviors. Along these lines, a 'mere exposure effect' was proposed that simply being exposed to a message could change attitudes and even behavior [6]. Research since then has tried to specify the conditions under which this might be true. Short (5 ms) exposures produced more stimuli recognition than longer exposures (500 ms) [7]. Repeated subliminal exposures (five) produced more positive effect responses to stimuli [8]. Asking about a person's intention to engage in a behavior increased the likelihood of performance of the behavior, including the purchase of big-ticket items such as automobiles [9–11]. The conversion

doi:10.1093/her/cyl087

from intent to behavior was more likely to occur when the participant did mental homework and could more easily represent or imagine doing that behavior [12]. This literature has not differentiated between whether respondents experienced changes in the underlying dimension or their position on the scale item (i.e. their interpretation of the items on the dimension). The latter could suggest that change in item response occurred without any necessary change in the underlying behavior or attitudinal variable. If this happens, treatment group participants' understanding of the items in behavioral or psychosocial assessment tools would be different without indicating a difference in behavior due to the intervention. One of the purposes of a comparison group in a research study is to control for measurement effects, such as the effect of simple exposure to an instrument, to differentiate them from the intervention effect in the treatment group. Heretofore, only whole instrument differences noted in the control group have been noted and discussed, however. It may be that some of the effects of mere exposure relate to specific items rather than the summary score of a whole instrument.

Differential item functioning (DIF) is a method that educational researchers have used to examine bias at the item level [13]. DIF analyses have been conducted most often to identify items that function differently among subgroups of participants. Specifically, these methods have been used to determine if participants who have similar scores on a given construct have the same interpretation of specific items. DIF analyses consist of focusing on the individual items themselves as a method for locating which items function differently between the two groups. DIF can determine if the properties of the items have changed as a result of participating in an intervention or from repeating the measurement. Item response modeling (IRM) [14] is one of the methods used to assess DIF. The properties of IRM, including stability of item and test parameters [14], make this a powerful DIF method.

To demonstrate the concept, this paper reports on two simple applications of a DIF analysis of self-reported behaviors (consumption of fruit and vegetables) and related psychosocial variables (preferences and self-efficacy). IRM DIF analysis was used to compare treatment and control groups after an intervention. The effect of taking the scales twice in the control group was also examined. Participant outcomes were reassessed after correcting for DIF.

Methods

Design

Two interventions with Boy Scouts were conducted: one to increase their fruit and vegetable (FV) consumption (5ADay Badge program) [15] and the other to increase their physical activity (Fit for Life Badge program) [16]. Forty-two troops were assigned randomly to intervention groups. Each treatment group served as the attention placebo control for the other, thereby enabling all troops to receive an intervention, and providing meaningful attention to each control group. Baseline data were collected at the troop meeting the week before the start of the intervention, the ninth week of the intervention (right after the badge awards ceremony) and 6 months after that. This report includes the baseline and immediate post-data relevant to the 5ADay Badge program only.

Intervention

The 5ADay Badge program was implemented across 9 weeks. Each week included an ~30-min in-troop educational experience and an ~20-min Internet experience. Paid badge troop leaders were trained in the delivery of the in-troop experience to ensure fidelity of delivery. The troop experience involved activities requiring personal contact: e.g. weekly recipe preparation (skill building) and taste testing (multiple exposures). FV recipes were provided in a Boy Scout cookbook and were previously tested to be enjoyed by many children/adolescents. Fidelity of the Internet program to desired theoretically specified procedures was assured by the computer programming. The Internet experience involved a weekly comic strip of Boy Scouts facing and overcoming challenges (modeling of problem solving) to eating FV in situations from those reported to be common problems

among children (thereby enhancing self-efficacy), anticipatory problem solving, personal goal setting, self-monitoring, goal attainment assessment, problem solving if goal not attained, point assignment (reward) and functional knowledge games (e.g. what counts as a fruit?). Logging on to the Internet component of the intervention was monitored by the paid troop leader and Scouts were encouraged to login if they had not. Badges were earned if 70% of maximum points were attained. The 5ADay Badge program has been described in greater detail elsewhere [15]. Analyses were performed at the individual scout level for simplicity, and because the same Internet program and trained badge troop leaders delivered all instruction.

Measures

Three FV-related measures were employed: frequency of consumption, preference and self-efficacy. Both the frequency and preference measures combined both FV items. The frequency of FV consumption measure (FvFreq) had 38 items. Participants responded with the categorized frequency of consumption of each item in the last week, with seven possible response categories ranging from 'none' to '15 or more servings last week'. We expect that these items assess a latent variable of perceived frequency of consumption of fruit and vegetables, with some items more frequently consumed by most people, even low consumers, and other items only consumed by more frequent consumers. Within the context of dietary change intervention, frequency of consumption is the primary dependent variable, and a DIF analysis would verify whether change in consumption of items occurred along this latent continuum, or is more random. The FV preference measure (FvPref) had 39 items, with three possible response categories, ranging from 'I do not like this' to 'I like this a lot'. We expect that these items assess a latent variable of preferences for fruits and vegetables with some items more commonly preferred by most people, even those not generally liking FV, and other items only preferred by those who generally prefer FV. The self-efficacy for eating FV measure (FvSE) had 21 items, with five possible response categories, ranging from 'disagree a lot' to 'agree a lot'. The preference and self-efficacy scales were previously construct validated using classical test theory procedures [5]. Cronbach's alpha varied from 0.72 to 0.87 [5]. As determinants of FV consumption, FV preferences and self-efficacy should be mediators of behavior change (but those analyses will not be reported here). Demographic characteristics were assessed with commonly used questions.

Psychometric analyses

The same procedures were used to analyze all three scales in ConQuest software [17]. IRM procedures for analyzing a scale have been described in detail elsewhere [14] and are only briefly summarized here.

Models were selected from the one-parameter Rasch family of item response models in order to estimate locations for both participants and items on the same scale. Pre-intervention and post-intervention data were analyzed first with a rating scale model and then with a partial-credit model [18] to determine whether the categories of responses were the same or different across items. The partial-credit model fit better for all measures, so for each item in each measure, the ordered categories of response (e.g. five ordered responses from disagree a lot to agree a lot) were modeled to have different values. The units for the relationship between participant characteristic and the endorsability of an item are expressed in logits (log of the odds) because the IRM procedures examine the probability of a particular response to an item.

To detect DIF, we introduce an interaction term in the model to estimate the probability of interaction between the item and the grouping variable (either treatment group or time in this study). The result of each DIF analysis is a list of estimated locations θ for the participants (each individual θ and their standard errors), a list of estimated locations δ for the items i and categories k (δ_{ik}) with their standard errors, and a list of estimated 'additions' γ (with their standard errors) to the item locations for each group g that correct for differences in the item functioning between the groups: responses in logits = $\theta - (\delta_{ik} + g_{ig})$. By examining

the g_{ig} additions and comparing them to their standard errors (e), we can see whether the DIF is statistically significant ($g_{ig} > 1.96e$). Since the average of the DIF effects is zero by definition, the overall difference in the locations of the participants should remain unchanged when comparing an analysis modeling DIF with a similar analysis without modeling DIF. We can check this observation by comparing the main effect of the grouping variable in the different analyses.

To examine item fit within the partial credit models used for these scales, ConQuest compared the observed item estimates with estimates that might have occurred if the model fit exactly, using a weighted fit and t statistic [17]. If both these values were outside the designated ranges (0.75–1.34 for weighted fit and -2 to $+2$ for t values), the item was considered misfitting. Items or categories that did not fit the model well were examined for possible substantive reasons. All analyses constrained mean item estimates to equal zero, and designated that ConQuest calculate full standard errors for items and steps from one category to the next.

For the initial assessment of participant outcomes, performed without accounting for DIF, we obtained group effects from pre- and post-intervention analyses for each of the three scales. The difference between the treatment and control groups' pre- and post-intervention was calculated as the difference between the mean estimated locations of each group. A difference was considered statistically significant (at $\alpha = 0.05$) if the distance from zero to either group mean was >1.96 times the standard error of the estimate. We hoped to see no significant difference between the treatment and control groups' pre-intervention since scout troops were randomly assigned to these groups. We also hoped to see a significant difference in the group effect post-intervention for all the three scales to reflect changes in responses attributable to the intervention. Significant differences in the group effects would indicate 'differential impact', a term used to indicate that the groups have different amounts of the underlying trait. We contrasted this term with DIF which indicates a difference in the way a particular item functions for members of different groups having 'the same amount' of the underlying trait.

In the DIF analysis, we compared post-intervention data from the treatment and control groups to see which, if any, items exhibited significant DIF following the intervention. Data were coded so that if g_{ig} for the treatment group was a positive number, then that item was more easily endorsed by the control group; if g_{ig} for the treatment group was a negative number, then that item was more easily endorsed by the treatment group.

The effect size of DIF was considered small if two times the g_{ig} was <0.426 logits, and moderate if it was between 0.426 and 0.638 logits. DIF >0.638 logits was considered large [19, 20]. Main effect of the grouping variable was obtained through the analysis that accounted for DIF. The main effect of group at post-intervention was expected to show significance for all three scales to reflect changes in the treatment group following intervention.

In the second analysis, pre- and post-intervention data (Time 1 and Time 2, 9 weeks apart, but no nutrition intervention) from the control group only were each calibrated to see which, if any, items exhibited significant difference following mere double exposure to the scales. Since the control group takes the items twice in these comparisons, jeopardizing the assumption of item independence needed to perform a DIF analysis as above, we used a method of comparing estimates from different occasions described by Wright and Masters [18]. The separately calibrated item estimates from pre- and post-intervention were compared using their means and variances and z-scores to set the pre-intervention estimates on the same post-intervention scale: $\delta'_{\mathrm{pre}i} = z_{\mathrm{pre}i}\,(\sigma_{\mathrm{post}i}) + \mu_{\mathrm{post}i}$; where $z_{\mathrm{pre}i} = (\delta_{\mathrm{pre}i} - \mu_{\mathrm{pre}i})/\sigma_{\mathrm{pre}i}$, $\mu_{\mathrm{pre}i}$ and $\mu_{\mathrm{post}i}$ are the pre- and post-intervention item calibration means, $\sigma_{\mathrm{pre}i}$ and $\sigma_{\mathrm{post}i}$ are the pre- and post-intervention item calibration standard deviations and $\delta_{\mathrm{pre}i}$ and $\delta'_{\mathrm{pre}i}$ are the pre-adjusted and post-adjusted item estimates for pre-intervention items. Then the standard errors (e_{12}) were calculated: $e_{12} = (e_1{}^2 + e_2{}^2)^{1/2}$. By comparing the differences with their standard errors, we judged whether they were statistically significant ($\delta_{\mathrm{pre}i} - \delta_{\mathrm{post}i} > 1.96e_{12}$).

Results

Demographics

Data were available on up to 473 Boy Scouts (Table I). There were no statistically significant differences between treatment and control groups for age, height, weight, body mass index (BMI), ethnicity, educational attainment or family owning a home. The data set included complete data on 228 participants in the treatment group and 230 in the control group.

Three measures were analyzed. Mean raw scores on each of the three measures are shown at the bottom of Table II with the maximum value of each measure appearing next to the measure name. There were no responses for FvFreq in item categories 5 and 7 for item 37 (cabbage); so the categories for this item were collapsed to make five categories.

As a precursor to the DIF analyses, we assessed the fit statistics of the items for their respective scales. Only a small number of items did not meet the fit criteria for these scales. In the FV frequency scale, consumption of 'French fries' was misfitting (weighted mean square = 1.37, t = 6.2) suggesting that responses to this item were more random than expected for this model. The item was retained because of its unique role as a putative vegetable, but one whose consumption should have been minimized on the basis of the intervention. In the self-efficacy scale, both eating fruit and eating vegetables served in school lunches were misfitting (weighted mean squares = 1.43 and 1.50, respectively, t = 8.0 and 9.3, respectively), again suggesting that responses to these items were more random than expected. These items were retained because of the prevalence of school lunches in the diets of these participants.

Table I. *Means (M), standard deviations (SDs), frequencies* (n) *and percentages (%) for baseline demographic characteristics and FV frequency, preference and self-efficacy by intervention group*

Variable	5ADay		Control		Total	
	n	M (SD)	*n*	M (SD)	*n*	M (SD)
Age (years)	229	12.8 (1.1)	238	12.8 (1.1)	467	12.8 (1.1)
Height (cm)	225	156.0 (11.2)	228	155.2 (10.0)	453	155.6 (10.6)
Weight (kg)	227	52.9 (17.2)	231	51.5 (13.2)	458	52.2 (15.3)
BMI (kg m^{-2})	224	21.4 (5.0)	227	21.1 (4.1)	451	21.3 (4.5)
BMI percentile	221	64.9 (28.8)	225	66.7 (29.1)	446	65.8 (29)
	n	%	*n*	%	*n*	%
Ethnicity						
Anglo-American	165	71.4	180	75.0	345	73.2
African American	11	4.8	6	2.5	17	3.6
Hispanic	29	12.6	35	14.6	64	13.6
Mixed/other	26	11.3	19	7.9	45	9.6
Total	231	100.0	240	100.0	471	100.0
Highest education in household						
High School graduate or less	10	4.3	17	7.2	27	5.8
Some college/technical school	50	21.7	60	25.3	110	23.6
College graduate	91	39.6	76	32.1	167	35.8
Post graduate	79	34.3	84	35.4	163	34.9
Total	230	100.0	237	100.0	467	100.0
Own home						
No	14	6.0	19	7.9	33	7.0
Yes	219	94.0	221	92.1	440	93.0
Total	233	100.0	240	100.0	473	100.0

Table II. *Mean raw scores for treatment and control groups pre- and post-intervention for FV frequency, preferences and self-efficacy*

Mean raw scores	Treatment group	Control group
FvFreq (maximum 264)		
Pre	67.2	64.5
Post	76.3	70.7
FvPref (maximum 78)		
Pre	41.6	39.5
Post	43.6	41.0
FvSE (maximum 84)		
Pre	63.3	60.8
Post	66.3	60.1

Tables III–V show the item estimates resulting from the FvFreq, FvPref and FvSE measure DIF analyses, respectively. Only items with statistically significant DIF have their DIF estimates listed under the DIF column. Although the DIF g_{ig} for the other items are not listed, the reader should keep in mind that all items have either a positive or a negative DIF g_{ig}, whether or not the magnitude is statistically significant, and that the sum of the positive g_{ig} must equal the sum of the negative g_{ig} in this model. A net positive or negative sum would indicate differential impact rather than DIF. For FvFreq, eight items showed significant DIF when analyzing treatment and control groups' post-intervention; all DIF were small and they were evenly distributed with four items showing positive DIF (grapes, apples, French fries and cantaloupe: the control group consumed these more frequently) and four items showing negative DIF (orange juice, carrots, oranges and pears: the treatment group consumed these more frequently). For FvPref, DIF for French fries was statistically significant and moderate in size, with the control group reporting more liking for them than the treatment group post-intervention. The control group also expressed more liking for watermelon (small DIF), whereas the treatment group expressed more liking for spinach, strawberries and lettuce post-intervention. For FvSE at post-intervention, the treatment group was more certain that they could eat fruit with a fruit dip, eat vegetables with a dip (both small DIF) or make a favorite fruit, juice or vegetable recipe (moderate DIF) than the control group.

Table VI shows the main effect of treatment group both pre- and post-intervention without accounting for DIF. Higher mean logit values indicated more of the underlying trait. The differences between treatment and control group main effects were significant ($P < 0.05$) for both FvFreq and FvSE post-intervention. No other main effect differences were significant. Main effects of the grouping variables were also estimated in the DIF analyses: treatment versus control group post-intervention (bottom of Table VI). After controlling for DIF, group membership was significant as a main effect in the frequency and self-efficacy analyses post-intervention, but not the preference analyses.

Mere double exposure to the three scales by the control group at Times 1 and 2 resulted in a significant difference between item calibrations for only one item in each of the three scales. For FvFreq, strawberries were consumed more frequently at Time 1 (-0.33 logits, $e_{12} = 0.15$, $P < 0.05$); for FvPref, 'cabbage' was better liked at Time 2 (0.47 logits, $e_{12} = 0.17$, $P < 0.01$) and for FvSE, 'I am sure I can eat 3 or more servings of vegetables every day' was more easily endorsed at Time 2 (0.40 logits, $e_{12} = 0.18$, $P < 0.05$).

Discussion

Analysis of individual items (Table III) showed that eight of the 38 FvFreq items (21%) demonstrated DIF between treatment and control groups' post-intervention. The number of items with statistically significant positive and negative g_{ig} was evenly balanced with four items each. The DIF effects were statistically significant, but deemed negligible according to DIF standards of effect size [19]. For treatment and control groups at post-assessment, three of the 21 FvSE items (14%), and five of the 39 FvPref items (13%) showed significant DIF. Thus, DIF was found in more items of the self-report of FV intake, compared with the number of

Table III. *FV frequency items with their logit value and estimate of DIF γ_{ig} where significant*

Items in order from most to least frequently consumed	Item parameters (post-intervention)	Standard errors	DIF[a] (treatment–control)	Standard errors
Other 100% juice	−0.519	0.045		
100% orange juice	−0.496	0.054	−0.145	0.038
Lettuce	−0.428	0.050		
French fried potatoes	−0.384	0.051	0.142	0.039
Greens (spinach, collard, turnip)	−0.336	0.052		
Bananas	−0.321	0.056		
Apples	−0.32	0.063	0.094	0.039
Grapes	−0.303	0.054	0.088	0.038
Carrots	−0.29	0.055	−0.142	0.040
Cooked beans (pinto, black-eyed beans)	−0.271	0.056		
Corn	−0.2	0.071		
Strawberry	−0.13	0.066		
Broccoli	−0.092	0.067		
Oranges	−0.089	0.071	−0.094	0.043
Apple sauce	−0.084	0.067		
Fruit salad or fruit cocktail	−0.081	0.064		
Green beans	−0.077	0.076		
100% apple juice	−0.076	0.076		
Tomatoes	−0.009	0.083		
Watermelon	0.011	0.068		
Celery	0.047	0.077		
100% grape juice	0.067	0.084		
Green peas	0.079	0.093		
Peaches	0.097	0.084		
Other white potato	0.113	0.108		
Cole slaw	0.154	0.079		
Grapefruit	0.169	0.079		
Pineapple	0.188	0.108		
Potato salad	0.193	0.094		
Pear	0.221	0.108	−0.140	0.051
Cantaloupe or muskmelon	0.242	0.108	0.133	0.050
Sweet potatoes	0.257	0.109		
Plums	0.322	0.110		
Raisins	0.325	0.134		
Kiwi	0.373	0.136		
Cabbage	0.379	0.129		
Okra	0.565	0.208		
Dried fruit	0.709	0.253		

[a]Positive γ_{ig} means the control group has said that they consume this item more frequently; negative γ_{ig} means the treatment group has said that they consume this item more frequently.

items showing DIF in the psychosocial correlates of FV intake, but all the intake DIFs were considered small, and thereby negligible.

The treatment–control group DIF could indicate any of several things: random variations in item functioning; artifact of the model and the need for DIF estimates to sum to zero; true change in frequency of consumption (which was reflected in differences between treatment and control at post) or difference in understanding items in relation to the underlying measure as a mere exposure effect. Given that 21% of the FVFreq items demonstrated

Table IV. *FV preference items with their logit value and estimate of DIF where significant*

Items in order of liking from most to least liked items	Item parameters (post-intervention)	Standard errors	DIF 1[a] (treatment–control)	Standard errors
Apples	−1.117	0.098		
Grapes	−1.04	0.091		
Corn	−1.023	0.09		
100% orange juice	−0.91	0.086		
Other 100% juice	−0.864	0.091		
French fried potatoes	−0.791	0.081	0.264	0.075
Strawberry	−0.747	0.075	−0.151	0.072
100% apple juice	−0.691	0.078		
Oranges	−0.676	0.077		
Watermelon	−0.661	0.073	0.151	0.07
Bananas	−0.501	0.073		
Carrots	−0.321	0.071		
Apple sauce	−0.266	0.066		
Pineapple	−0.199	0.065		
100% grape juice	−0.169	0.068		
Lettuce	−0.105	0.067	−0.148	0.067
Peaches	−0.011	0.065		
Pear	0.014	0.065		
Other white potato	0.064	0.065		
Brocolli	0.102	0.065		
Fruit salad or fruit cocktail	0.127	0.064		
Greens	0.146	0.068		
Cooked beans (pinto, black-eyed beans)	0.16	0.064		
Green beans	0.194	0.065		
Cantalope or muskmelon	0.266	0.066		
Kiwi	0.279	0.064		
Tomatoes	0.332	0.064		
Plums	0.339	0.064		
Green peas	0.353	0.066		
Celery	0.411	0.065		
Raisins	0.557	0.069		
Grapefruit	0.697	0.067		
Spinach	0.729	0.068	−0.136	0.067
Potato salad	0.772	0.069		
Dried fruit	0.842	0.072		
Cabbage	0.849	0.071		
Sweet potatoes	0.871	0.068		
Cole slaw	0.98	0.069		
Okra	1.011	0.074		

[a]Positive numbers mean these items were more liked by the control group; negative numbers mean these items were more liked by the treatment group.

DIF, it seems unlikely that all of the observed DIF would be due to chance. The underlying measure reflects the self-reported frequency of consumption of the items, with items at the negative end of the logit scale reflecting more frequent consumption. The fact that the treatment group consumed some items more frequently than the control group after receiving an intervention that promoted increased consumption of all these items is not unexpected if the DIF reflects true changes in intake. The

Table V. *FV self-efficacy items with their logit value and estimate of DIF where significant*

Items in order of agreement from most to least readily agreed upon statements	Item parameters (post-intervention)	Standard errors	DIF 1[a] (treatment–control)	Standard errors
I am sure I can eat my favorite fruit with lunch at home on the weekends	−0.418	0.088		
I am sure I can ask someone in my family to buy my favorite fruit or vegetables	−0.379	0.083		
I am sure I can go shopping with my family for my favorite fruit or vegetables	−0.344	0.079		
I am sure I can drink a glass of my favorite 100% fruit juice for breakfast	−0.296	0.075		
I am sure I can ask someone in my family to make my favorite vegetables for dinner	−0.249	0.075		
I am sure I can eat fruit or vegetables for lunch in front of my friends	−0.214	0.074		
I am sure I can eat my favorite vegetables for lunch at home	−0.205	0.079		
I am sure I can eat a fruit or vegetable for dinner when I eat out or away from home	−0.135	0.078		
I am sure I can eat 2 or more servings of fruit every day	−0.114	0.072		
I am sure I can prepare my favorite fruit or vegetable to eat.	−0.106	0.068		
I am sure I can add fruit or vegetables to my normal school lunch	−0.033	0.064		
I am sure I can make my own dinner that includes a fruit or vegetable when someone else doesn't have time to cook	−0.018	0.069		
I am sure I can eat 3 or more servings of vegetables every day	0.031	0.073		
I am sure I can make my favorite fruit, juice and vegetable recipes	0.133	0.064	−0.267	0.063
I am sure I can add my favorite fruit to my cereal at breakfast.	0.138	0.06		
I am sure I can add my favorite vegetables to my favorite sandwich for lunch at home	0.144	0.058		
I am sure I can eat my favorite fruit with a fruit dip for a snack	0.201	0.061	−0.123	0.056
I am sure I can eat my favorite raw vegetables with dip for a snack	0.326	0.057	−0.164	0.054
I am sure I can eat my favorite fruit instead of my usual dessert	0.369	0.056		
I am sure I can eat a fruit that's served at school lunch	0.507	0.053		
I am sure I can eat a vegetable that's served at school lunch	0.66	0.053		

[a]Positive numbers mean these items were more easily endorsed by the control group; negative numbers mean these items were more easily endorsed by the treatment group.

Table VI. *Group effect (in logits) both pre- and post-intervention without and with accounting for DIF*

	FvFreq		FvPref		FvSE	
	Treatment	Control	Treatment	Control	Treatment	Control
Not accounting for DIF						
Pre-intervention						
Mean (standard error)	0.024 (0.022)	−0.024	0.061 (0.034)	−0.061	0.053 (0.060)	−0.053
Estimated difference	0.048		0.122		0.106	
Post-intervention						
Mean (standard error)	0.136 (0.018)	−0.136	0.055 (0.032)	−0.055	0.297 (0.025)	−0.297
Estimated difference	0.272[a]		0.11		0.594[a]	
After accounting for DIF						
Mean (standard error) in logits	0.134 (0.019)	−0.134	0.056 (0.032)	−0.056	0.302 (0.025)	−0.302
Difference between mean estimates in logits	0.268[a]		0.112		0.604[a]	

Higher logit values meant participants had more of the underlying trait: frequency of consumption, liking or self-efficacy.
[a]Group effect was significant.

fact that the control group consumed some items more frequently than the treatment group is less explainable, although French fries were specifically addressed in the intervention as less desirable due to high fat and thus perhaps less frequently consumed by the treatment group post-intervention.

Future intervention studies should combine self-report methods and more objective markers (e.g. blood markers), and employ IRM and DIF techniques to correct for this problem, thereby obtaining a clearer assessment of the functioning of different items in comparison to the more objective markers. The greater preference for French fries at post-assessment in the control group appears to reflect an intervention effect, since scouts in the intervention group were clearly informed that French fries did not count as a vegetable. The moderate increase in self-efficacy for making vegetable recipes should reflect the scout participation in weekly recipe preparation in troop meetings. The small effects for eating fruit and vegetables with dip should also reflect the intervention emphasis on these issues. Thus, the DIF appears to have identified change in a few aspects of frequency, preference and self-efficacy specifically changed by the intervention. Given the 'negligible' level of DIF on all the items, these analyses provide little support for a mere exposure effect on understanding of these items, but the possibility of mere exposure effect on the mean change in location (differential impact) on the latent variable for behavior and self-efficacy remains.

Table VI showed that statistically correcting for treatment–control group DIF at post-intervention resulted in no change in outcome main effects. Thus, DIF did not mask differential impact.

In this paper, we assessed possible DIF as an artifact of taking the measure twice by assessing possible differences in item calibrations from two different occasions of taking the scales. A total of three items showed different calibrations from Time 1 to Time 2. One item in each of the FvFreq and FvSE scales showed a DIF analogue at the 95% confidence level, about what might be expected by chance for scales of $\geqslant$20 items. Thus, the evidence is weak for mere exposure effect differences in how these Boy Scouts responded to these two scales on a second occasion. The evidence is slightly stronger for a mere exposure effect on the FvPref scale because the DIF analogue is larger for liking for cabbage, significant at the 99% confidence level for this 39-item scale. Further research might confirm this finding and clarify its interpretation.

These analyses were kept as straightforward as possible to demonstrate the DIF analysis procedures within IRM. A better approach to determining differences in item calibrations from different occasions would be to assess 'person DIF' but

this analysis was deemed outside the scope of this paper. A more complex analysis that allows for simultaneous measurement, treatment and measurement by treatment DIF estimation is possible, but, again, was judged too complex for the purpose of demonstrating DIF procedures.

The strengths of this paper were the reasonably large sample, the use of measures that were previously validated using classical test theory procedures and application of sophisticated psychometric analysis procedures. The weaknesses include that these measures were not previously validated using item response theory methods, that the control group received a physical activity change intervention which may have confounded the 'mere measurement effect', and no hierarchical analysis was performed to account for the clustering of data by troop.

Conclusion

These analyses demonstrated that some items functioned differently after participants participated in an intervention and after some were exposed to a measurement scale but these effects were mostly small and could be considered negligible. This may be problematic for the usual interpretation of item functioning. Future research will need to replicate these results.

Acknowledgements

This research was supported by a grant from the American Cancer Society (ACS TURSG-01) and a contract from the National Cancer Institute (NCI 263-MQ-31958). This work is also a publication of the USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston, TX, USA. This project has been funded in part by federal funds from the United States Department of Agriculture/Agricultural Research Service (USDA/ARS) under co-operative agreement 58-6250-6001. The contents of this publication do not necessarily reflect the views or polices of the USDA or NCI, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

Conflict of interest statement

None declared.

References

1. Baranowski T, Lin LS, Wetter DW *et al.* Theory as mediating variables: why aren't community interventions working as desired? *Ann Epidemiol* 1997; **7**: S89–95.
2. Baranowski T, Anderson C, Carmack C. Mediating variable framework in physical activity interventions. How are we doing? How might we do better? *Am J Prev Med* 1998; **15**: 266–97.
3. Baranowski T, Jago R. Understanding mechanisms of change in children's physical activity programs. *Exerc Sport Sci Rev* 2005; **33**: 163–68.
4. Domel SB, Baranowski T, Davis H *et al.* Fruit and vegetable preferences among fourth and fifth grade students. *Prev Med* 1994; **22**: 866–79.
5. Domel SB, Baranowski T, Thompson WO *et al.* Psychosocial predictors of fruit and vegetable consumption among elementary school children. *Health Educ Res Theory Pract* 1996; **11**: 299–308.
6. Zajonc RB. Attitudinal effects of mere exposure. *J Pers Soc Psychol* 1968; **8**: 1–29.
7. Bornstein RF, D'Agostino PR. Stimulus recognition and the mere exposure effect. *J Pers Soc Psychol* 1992; **63**: 545–52.
8. Monahan JL, Murphy ST, Zajonc RB. Subliminal mere exposure: specific, general, and diffuse effects. *Psychol Sci* 2000; **11**: 462–6.
9. Sherman SJ. On the self-erasing nature of errors of prediction. *J Pers Soc Psychol* 1980; **39**: 211–21.
10. Greenwald AG, Carnot CG, Beach R *et al.* Increasing voting-behavior by asking people if they expect to vote. *J Appl Psychol* 1987; **72**: 315–8.
11. Morwitz VG, Johnson E, Schmittlein D. Does measuring intent change behavior. *J Consum Res* 1993; **20**: 46–61.
12. Levav J, Fitzsimons GJ. When questions change behavior: the role of ease of representation. *Psychol Sci* 2006; **17**: 207–13.
13. Shepard LA. Definition of bias. In: *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press, 1982.
14. Wilson M. *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Erlbaum, 2005.
15. Thompson D, Baranowski T, Cullen K *et al.* 5 A Day Boy Scout Badge program: outcome results. *Am J Prev Med* 2006, in press.

16. Jago R, Baranowski T, Baranowski J *et al*. Fit For Life Boy Scout badge: outcome evaluation of a troop & Internet intervention. *Prev Med* 2006; **42**: 181–7.
17. Wu ML, Adams RJ, Wilson MR. *ACER ConQuest: Generalised Item Response Modelling Software*. American College of Educational Research, 1998.
18. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: Mesa Press, 1982.
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. Longford N, Holland P, Thayer D. Stability of the MH D-DIF statistics across populations. In: Holland P, Wainer H (eds). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1993, 171–96.

Received on January 10, 2006; accepted on July 19, 2006