DOT/FAA/AM-01/4
Office of Aviation Medicine
Washington, D.C. 20591

Latent Trait Theory Analysis of Changes in Item Response Anchors

William L. Farmer
Richard C. Thompson
Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

Susan K.R. Heil
Oklahoma Department of Mental Health
& Substance Abuse Services
Oklahoma City, OK 73152

Michael C. Heil
Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

February 2001

Final Report

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

20010328 075

U.S. Department of Transportation
Federal Aviation Administration

N O T I C E

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-01/4		

4. Title and Subtitle	5. Report Date
Latent Trait Theory Analysis of Changes in Item Response Anchors	February 2001
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Farmer, W.L.[1], Thompson, R.C.[1], Heil, S.K.R.[2], & Heil, M.C.[1]	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
[1]FAA Civil Aeromedical Institute Oklahoma City, OK 73125 [2]Oklahoma Department of Mental Health & Substance Abuse Services Oklahoma City, OK 73152	
	11. Contract or Grant No.
12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S. W. Washington, D.C. 20591	
	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved subtask AM-B-99-HRR-516.

16. Abstract

The purpose of this study was to evaluate the effect that modifications in item response anchors have on responses to survey items. Twenty-nine items were administered in 1993 and 1995 as part of more extensive attitude surveys to two random samples of Federal Aviation Administration employees. Changes in the response scales (5-point Likert) between the two survey administrations ranged from no change at all to extensive re-anchoring of the response categories. Item responses were modeled via two-parameter graded response models based on item response theory. Changes in the way the item responses functioned between both years were assessed using the differential item functioning (DIF) method recommended by Muraki (1997). Twenty-four of the 29 items displayed significant levels of DIF, indicating that the response categories did not measure the constructs of interest in a similar fashion across the two administrations. Items whose response anchors had been changed substantially exhibited significant DIF more frequently than those where the change in anchors was less drastic. These results suggest that researchers and practitioners take a conservative approach when considering the revision of measuring scales for a particular set of items.

17. Key Words	18. Distribution Statement
Item Anchors, Latent Trait Theory, Item Response Theory, Differential Item Functioning, Measurement Equivalence	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	18	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Latent Trait Theory Analysis of Changes in Item Response Anchors

Landy, Shankster, and Kohler (1994) referred to the history of personnel selection as "pre-Copernican," in that the employer's goals and needs were put at the center of the universe with the value of other perspectives (i.e., applicants) considered only as a secondary concern. Similarly, in regard to attitude measurement, Schwarz and Hippler (1987) alluded to the fact that questionnaire developers/administrators have tended to view items and response categories presented as reactive measuring devices with negligible effect on actual responses observed.

Focusing on response anchors presented as item alternatives, Chang (1997) pointed out that authors often do not disclose when behavioral/attitudinal anchors are modified. The assumption that follows is that slight changes in rating scale anchors are unimportant. This assumption runs counter to the existing plethora of advice and "rules-of-thumb" regarding item construction and other psychometric issues.

Numerous studies have examined the effects of various characteristics of response scales. Among these have been reports addressing the optimal number of response categories (Alwin, 1997), the distributional and assumed interval-level qualities of scale scores (Dobson & Mothersill, 1979; Hofacker, 1984; Schriesheim & Schreisheim, 1974), and the meaning of particular response options to respondents (DuBois & Burns, 1975; Pace & Friedlander, 1982). These studies reveal that response category formats are an important element in survey construction.

Schwarz and Hippler (1987) stated that respondents utilize the information in response categories, including phraseology, when making decisions about answering an item. Respondents are generally asked a question in reference to a "real situation;" however, the response that is obtained may actually be more a function of the question and answer categories than the scenario that they represent (Rockwood, Sangster, & Dillman, 1997). Schwarz and Hippler (1987) referred to the social information processing model (Boudenhausen & Wyer, 1987) as an explanation for the effect that response categories have on the processing of survey items. Based on this model, respondents may use the response categories to assess the meaning of an item, as a frame of reference in estimating behavior frequency or quantity, or in determining via social comparison what is typical (Rockwood et al., 1997).

Studies have focused on the effect of response categories and the interaction between the types of questions to which the categories are linked and the mode of survey administration. For example, Rockwood, Sangster, and Dillman (1997) found that, when the target behavior/attitude is not well defined, frequent and mundane questions were more sensitive to changes in the response categories. They also found that the mode of administration (telephone interview vs. mail) interacted with variations in response categories to affect responses.

Whereas it is generally accepted that response category anchors are important, the degree of importance is unclear. Chang (1997) referred to a number of studies that present conflicting evidence about the effect of response category anchors on scale results. Based on the results of his own study, utilizing indices derived from generalizability theory, Chang concluded that "attitude measurement from a Likert-type scale can be generalized across different anchoring labels," and that future researchers could be less concerned with effects that might result from the use of different response anchors for Likert-type scales.

Do response anchors matter? The purpose of this study was to investigate the effect that response anchor modifications have on item functioning. In this study, the survey item results of respondents from two separate administrations of a large biennial employee attitude questionnaire were compared for differences that may have resulted from modifications to 29 items. Due to the desirable properties of invariant non-sample, specific item parameters, item response theory (IRT) provides an excellent vehicle for the evaluation of differential item functioning (DIF). Significant levels of DIF were expected for those items where response anchors differed between administrations. In addition, the amount of DIF was expected to be greater for those items where the changes in the scale anchors deviated more from their original format.

METHOD

Subjects

Employee attitude surveys that assessed a number of organizationally relevant variables (item *n*: 1993=146, 1995=138) were sent to representative stratified random samples of the FAA's workforce in 1993 (8,311 surveys sent) and 1995 (6,874 surveys sent). Each sample represented 15% of the total employee population for that year. A smaller number of surveys was sent in 1995, since there was no longer a requirement to sample employees involved in a pay demonstration project, as had been done in 1993. In addition, organizational downsizing, resulting from a voluntary employee buyout in 1994, contributed to the smaller sample in 1995. Return rates for the two administrations were 59% and 52% respectively, yielding a total of 8,393 returned surveys for the two years.

Survey Items

Between both administrations, 29 of the items were essentially identical in that they assessed the same domain (Appendix A). All 29 of the selected items were scaled via a 5-point Likert-type format. Modifications were made to these 29 survey items across a number of constructs (or dimensions). These included items assessing the impact of technology, communication effectiveness, training opportunities, employee empowerment, recognition and rewards, the perceptions of the organization's commitment to maintaining a positive and fair workplace, and customer support.

The extent of item match between the two versions ranged from exact duplicates, with regard to stem and response anchors, to those items whose anchors were changed from "agree/disagree" to "extent of . . agreement with." The types of changes made to items fell into four categories. These categories and the frequency of items reflecting the associated changes were a) Type 1 - identical response anchors (5 items), b) Type 2 - anchors where the mid-point was changed from "neither disagree nor agree" to "neutral" (10 items), c) Type 3 - response scales changed from "extent of" to "agree/disagree" (11 items), and d) Type 4 - response scales changed from "agree/disagree" to "extent of" (3 items). The exact scales used for each administration from each category are presented in Table 1 (see Appendix for complete listing of the 29 survey items).

Analyses

Item Response Theory (IRT). Two-parameter multiple-category IRT models of the 29 items were modeled utilizing the *PARSCALE 3.0* program (Muraki & Bock, 1997). Due to the graded response nature of the scales, Samejima's (1969) model with a logistic response function was used. Item-response models provide an invariant non-sample specific estimation of item difficulty or threshold, item discrimination, guessing, and, in multiple-category response data, category location parameters. The essence of item response theory (Lord, 1952) is that the probability of answering an item correctly or of attaining a particular response level is modeled as a function of an individual ability or latent trait. The most salient representation of this relationship is the item characteristic curve (ICC) (Figure 1). As an individual's ability or trait-level rises, the probability of answering an item correctly or at a specified-level, rises as well.

The two-parameter binary response model can be represented via the well-known formulation:

$$P(\theta)=\frac{\exp^{Da(\theta-b)}}{1+\exp^{Da(\theta-b)}}$$

where: θ = the latent construct of interest (being measured),
$P(\theta)$ = the probability of a specific response given θ,
D = a constant (most often 1.7),
a = the discrimination parameter of the item, and
b = the difficulty or threshold parameter of the item.

Conceptually, the b parameter can be thought of as the point on the latent construct scale θ where the probability of a given response is equal to 0.50. In more concise terms, it is the item's location parameter, and is analagous to the p-value (percentage of those responding correctly or at a prespecified level, for survey items) as an index of item difficulty in classical measurement theory. The a parameter represents the point on the latent trait scale where the item best discriminates between those of low or high levels of the construct of interest. Graphically, it is represented as the slope of the logistic function (i.e., a sigmoid cumulative curve) at the point of inflection. Its classical analog is the correlation between an item's score (represented on a test item by the proportion of respondents correctly answering the item and

Table 1
Item Response Anchors for Survey Items at Both Administrations

Change Type	*Response Category*	*Survey Administration* 1993	1995
1. Identical	1	Not at all	Not at all
	2	To a limited extent	To a limited extent
	3	To a moderate extent	To a moderate extent
	4	To a considerable extent	To a considerable extent
	5	To a very great extent	To a very great extent
2. Similar	1	Strongly disagree	Strongly disagree
	2	Disagree	Disagree
	3	Neither disagree nor agree	Neutral
	4	Agree	Agree
	5	Strongly agree	Strongly agree
3. "Extent of" To "Agree"	1	Not at all	Strongly disagree
	2	To a limited extent	Disagree
	3	To a moderate extent	Neutral
	4	To a considerable extent	Agree
	5	To a very great extent	Strongly agree
4. "Agree" to "Extent of"	1	Strongly disagree	Not at all
	2	Disagree	To a limited extent
	3	Neither Disagree nor agree	To a moderate extent
	4	Agree	To a considerable extent
	5	Strongly agree	To a very great extent

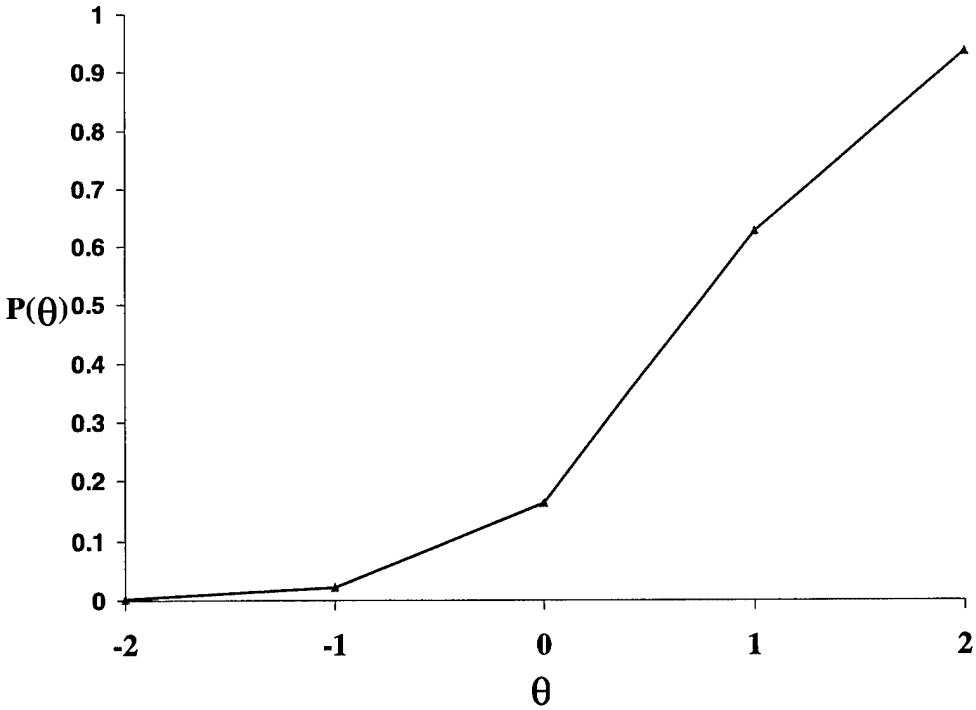

Figure 1. Item characteristic curve (ICC) for binary-response item.

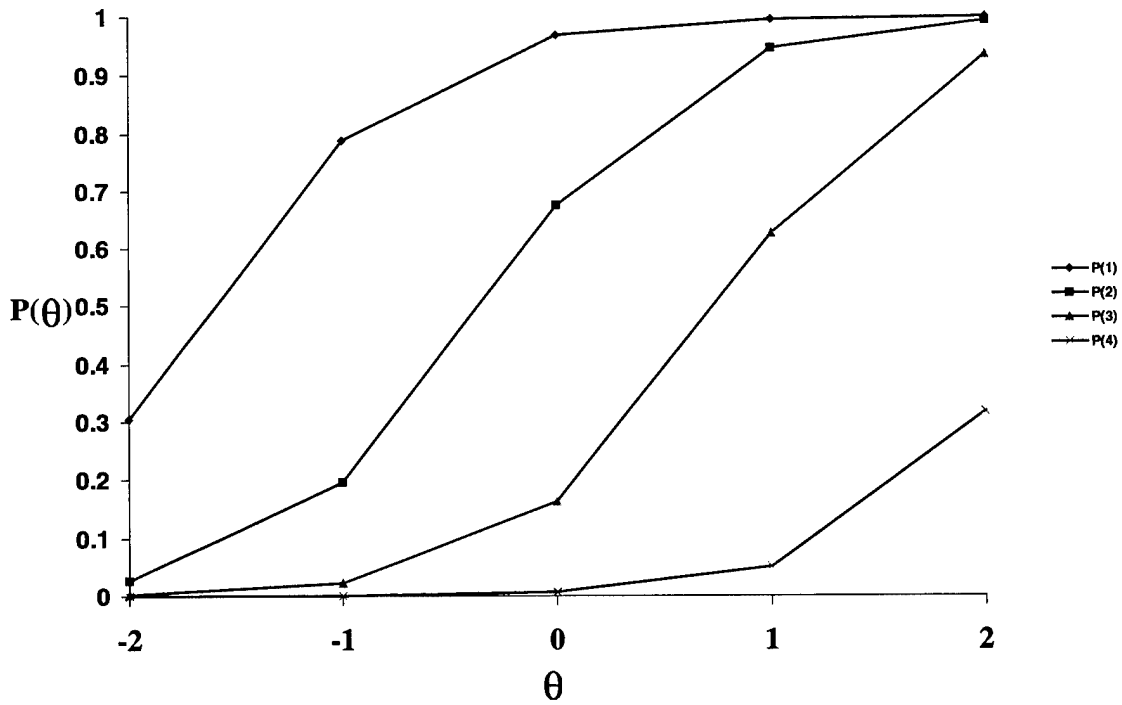

Figure 2. Item characteristic curves (ICCs) for graded-response (5 possible response options) item. Each curve represents the probability of selecting a response category over the one immediately preceding it, along the response continuum.

on a survey item by the mean item score across respondents) and the total test or survey score, provided that the score represents some linear composite of the raw or weighted item responses (e.g., a summed total of the item scores).

Though originally developed around a binary response framework (i.e., right-wrong), there currently exist 30 different major IRT models, and numerous extensions of these, that are suitable for assessing test performance, attitude and personality measurement (via questionnaire data), physiological performance, and a number of other assessment scenarios that warrant mathematically-based measurement models (van der Linden & Hambleton, 1997). For the purpose of Likert-type survey data, Samejima (1969) developed a graded response model that ideally represents the pattern of responses that is characteristic of that item-type. In the graded response model, the ICC represents the probability of responding in a particular category versus answering in the next lower category (e.g., the probability of answering "very satisfied" as opposed to "satisfied" for an item that assesses satisfaction). In this case, the *b* parameter no longer represents "item difficulty," but category placement along the latent trait continuum. Since each item response function represents the probability of being in response category, as opposed to the next category, there is one fewer ICC than the number of possible response options.

The item response function, or now category response function, is represented by:

$$P_u(\theta) = P_u^*(\theta) - P_{(u+1)}^*(\theta)$$

where: $P_u(\theta)$ = the probability of choosing a specific response category given θ,
$P_u^*(\theta)$ = the conditional probability of choosing a specific response category given θ,
$P_{(u+1)}^*(\theta)$ = the conditional probability of choosing the next higher response category given θ.

An example of a 5-category, yielding 4 ICCs, is illustrated in Figure 2.

Differential Item Functioning (DIF). DIF refers to a psychometric difference in how an item functions for two groups (Dorans & Holland, 1993). This is to be distinguished from test bias, which relates the differential performance of two groups on *total* test scores. Items are matched on a particular construct being measured, and DIF exists when the two groups differ on item performance with respect to the construct in question. Here performance is not limited to the narrow concept of the proportion of those answering correctly, or in the case of survey items attaining a particular mean item score. Performance also includes the ability of the item to distinguish between those high or low in the amount of a particular construct (e.g., those higher in extroversion). For multiple category items, such as Likert-type items, the focus would be placed on the item's performance at distinguishing consistently the "amount" of a construct (or attitude that respondents possessed in relation to the measuring instrument). DIF is distinguishable from *item bias* in that the latter term is broader and carries with it social connotations regarding item fairness. Holland and Wainer (1993) point out that DIF is limited to technical or statistical considerations and should not be confused with item bias. DIF assessment is appropriate in all measurement situations where the question of measurement consistency is a concern. Recent studies utilizing DIF-methodology to assess the measurement quality of instruments geared toward attitude and value measurement include those by Gable & Wolf (1993) and Lynch, Barnes-Farrell, and Kulikowich (1998). It is important to remember that DIF-methodology is not geared toward assessing changes in mean scores, as one would expect in an organizational survey where certain attitudes were affected by some organizational intervention. DIF-methodology is aimed at an assessment of the stability of an item's precision across varying conditions. These conditions may be defined by demographic group membership, time intervals, or any other variable that may affect the ability of an item to measure in a consistent manner. Wright (1979) uses the analogy of comparing a survey item to a ruler. Across time an individual may grow in height, yielding a change in the person's position on a ruler. Across various time intervals or between different persons, the ruler will measure an inch the same way every time. If the ruler is made of rubber or if we define an inch arbitrarily, we have no consistent means for measuring length. Likewise, though an individual's attitude may change (as indicated by a greater amount of organizational satisfaction, for instance), we, as measurement specialists, want to know that we are assessing a particular construct the same way every time. This should be independent of anything that we may hypothesize, or

not, to impact the quantification of a construct. DIF-methodology is the state-of-the-art set of tools for measurement specialists to assess the stability of their "ruler" (i.e., survey or test items).

A number of procedures exist for the assessment of DIF and include those based on classical test theory, contingency table analysis (e.g., Mantel-Haenszel, 1959), and IRT. Provided that the necessary assumptions are met and that sample sizes are sufficient, methods based on IRT are preferable. DIF may exist whenever any of the item parameters of an IRT model are different for the two groups of interest. Prior to the advent of modern measurement theory, as represented by IRT, assessment of DIF was limited to ad hoc methods based on classical test theory that were subject to such problems as the confounding of item difficulty and discriminability (Camilli & Shepard, 1994). Camilli and Shepard (1994) point out that these older methods tended to provide results that were intimately tied to extremeness of item response proportions and total sample size, and did not provide reliable, generalizable results. Figure 3 presents an example of demonstrable DIF for a binary-response item. As can be seen, the response functions for this item are different for two groups. Figure 4 illustrates the generalization of the binary-case to a Likert-type item with 5 response categories.

Based on the advice of Muraki (E. Muraki, personal communication, October 27, 1997), DIF for this study was limited to a difference in the difficulty (*b*) parameters estimated for the two groups. The fundamental notion here is that, though the *a* parameters are unconstrained (allowing them to vary from item to item), they are held constant between the two groups. This differs from the popular Rasch model, where the *a* parameters are held constant between groups of respondents and also between items.

The procedure involves estimating the IRT models for each group separately, and comparing the *b* parameters via the formula:

$$SID = \frac{b_{group1} - b_{group2}}{Var\,b_{group1} + Var\,b_{group2}}$$

where: *SID* = the standard index of DIF (a *z-score*) for the tested item. Existence of DIF is indicated by a statistically significant *SID* ($\pm$ *1.96*, $p \leq .05$).

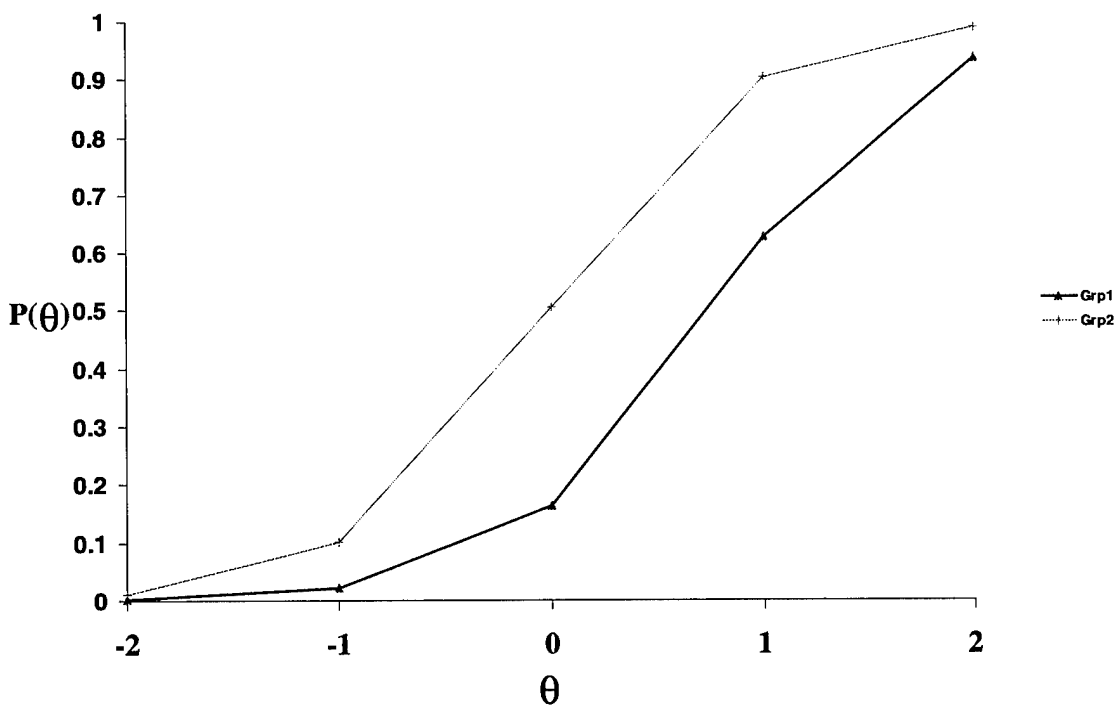

Figure 3. Binary-response item displaying measurable differential item functioning (DIF) between two groups. DIF is graphically displayed by the requirement to model separate curves based on group membership.

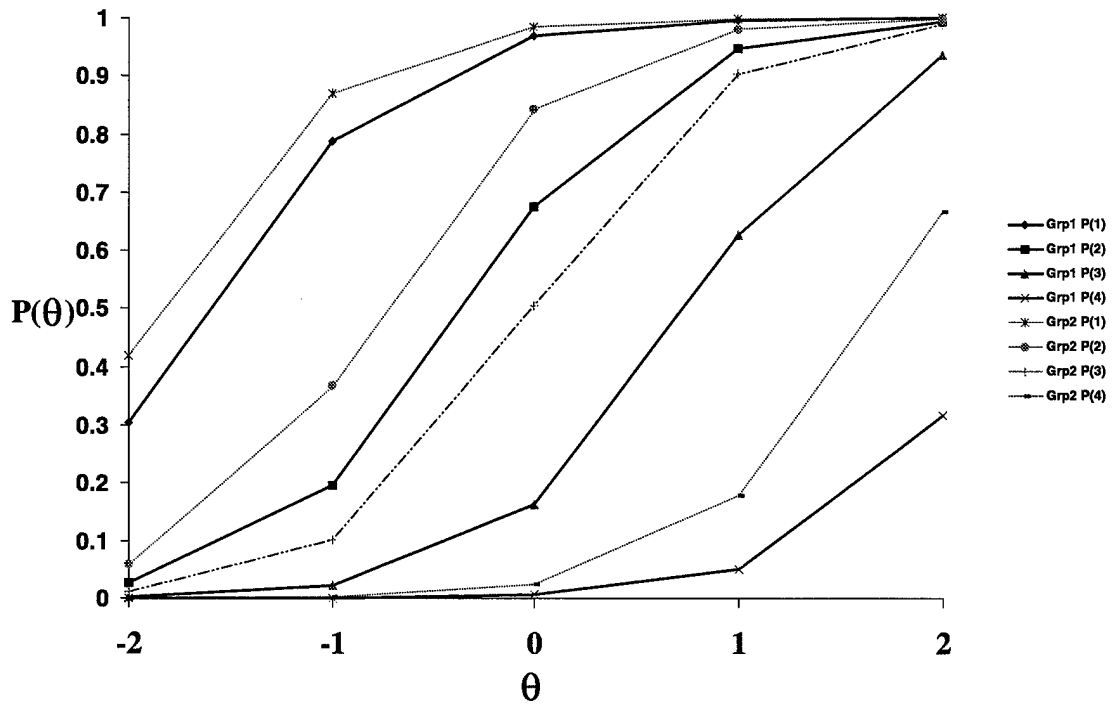

Figure 4. Graded-response item (5 options) displaying DIF.

RESULTS

Table 2 illustrates the results of t-tests of mean score differences on the 29 items compared between the 1993 and 1995 survey administrations. Significant differences existed between the two samples for all but three items. Effect size indices (*d*) were calculated according to the procedures specified by Cohen (1988, p. 20). Eighteen of the 29 obtained *d*'s failed to achieve a level indicative of a "small" effect size and, of the 11 that demonstrated an effect size, only one reached the "medium" level.

The results from the DIF analyses are presented in Table 3. DIF analyses revealed that 24 of the 29 items displayed statistically significant differential item functioning (see last column in table). Of the five items where no change (Type 1) had been made between the 1993 and 1995 administrations, three displayed DIF. Of note is that the items that did display DIF appeared to have fairly large values for the SID. Using the absolute value of the SIDs, the mean SID for the items that displayed significant DIF was 10.459. The mean SID value for all items combined was 6.591. As shown in Table 2, there was a substantial difference between those items that did not have significant SIDs and those that did have significant SIDs. This is displayed by the difference in mean SID for DIF items and for all items.

For Type 2 changes (similar item response anchors), seven of the ten selected items had statistically different *b* parameters between the two groups. The SIDs for Type 2 changes displayed a narrower range (as opposed to Type 1 changes) with a mean SID of 6.856 for DIF items and 5.125 for all items. All of the items (11 for Type 3, and 3 for Type 4) that had major changes to the scale response anchors displayed significant DIF indices. For the 11 Type 3 and three Type 4 change items, the mean SIDs were 13.989 and 15.849, respectively. Overall, the SIDs for Type 3 and Type 4 were quite large. In statistical terms, the obtained SIDs indicate that statistical significance is attributable to differences that really exist and are not merely an artifact of measurement error.

Items were less likely to differ on item parameters between groups when attitude anchors were more similar between years. This finding was consistent with the hypothesis that the amount of DIF would be greater for items where the anchor changes were more substantial. However, this finding counters the argument by some that "researchers need not be overly concerned with the practice of using different labels to anchor Likert-type scales for items of the same or different instruments" (Chang, 1997, p. 805).

Table 2
Results of t-tests Between 2 Survey Administrations for 29 Items

Change Type	Item	Means (93, 95)	SDs (93, 95)	*t*	df	*d* Effect *b*
1.	1	2.64, 2.52	0.99, 0.96	5.89*	8248	.12
	2	2.53, 2.40	0.97, 0.91	6.01*	7788.97*a*	.14
	3	2.49, 2.53	1.07, 1.00	-2.00*	7758.22 *a*	.04
	4	2.72, 2.98	1.15, 1.15	-9.84*	7564.81 *a*	.22
	5	2.39, 2.46	1.12, 1.10	-2.97*	8186	.06
2.	6	3.04, 2.77	1.10, 1.21	10.42*	7191.78 *a*	.24
	7	2.56, 2.64	1.09, 1.13	-3.01*	7485.37 *a*	.07
	8	3.02, 3.12	1.06, 1.14	-4.29*	7269.90 *a*	.09
	9	3.72, 3.42	1.09, 1.17	11.71*	7297.80 *a*	.27
	10	3.32, 3.15	1.18, 1.17	6.18*	8278	.14
	11	3.08, 3.11	1.16, 1.15	-1.04	7605.27 *a*	.03
	12	2.80, 2.73	1.19, 1.20	2.54*	7584.36 *a*	.06
	13	2.46, 2.29	1.15, 1.10	6.72*	7766.92 *a*	.15
	14	2.84, 2.68	1.17, 1.17	6.12*	7597.08 *a*	.14
	15	3.35, 3.37	1.02, 0.99	-1.00	7678.87 *a*	.02
3.	16	2.89, 3.20	1.17, 1.17	-11.96*	7631.34 *a*	.26
	17	3.25, 3.51	1.19, 1.13	-10.20*	7848.89 *a*	.22
	18	4.02, 4.17	0.95, 0.74	-8.02*	8294.05 *a*	.17
	19	2.51, 2.62	1.16, 1.17	-4.22*	7537.47 *a*	.10
	20	2.81, 2.98	0.97, 1.13	-7.20*	6903.18 *a*	.16
	21	2.42, 2.82	1.06, 1.11	-16.49*	7353.75 *a*	.36
	22	2.58, 3.01	1.14, 1.18	-16.84*	8253	.37
	23	2.58, 2.86	1.01, 1.17	-11.27*	6837.79 *a*	.26
	24	2.24, 2.63	1.16, 1.09	-15.29*	7744.38 *a*	.34
	25	2.20, 2.37	1.01, 1.09	-7.27*	7174.78 *a*	.16
	26	2.08, 2.71	0.89, 1.00	-29.94*	7060.26 *a*	.64
4.	27	3.64, 3.24	0.95, 1.10	17.20*	6854.75 *a*	.39
	28	3.35, 3.15	1.07, 1.04	8.80*	7671.60 *a*	.19
	29	3.00, 3.02	1.30, 1.65	-0.40	6591.72 *a*	.01

*Total N=8393 (1993 n=4825, 1995 n=3568); * - p ≤.05 (2-tailed);*
a - degrees of freedom adjusted for availability of sample and heterogeneity of variance;
b – small = .2, medium = .5, large = .8 (Cohen, 1988, p. 25).

Table 3
Estimated IRT parameters, Standard Errors of Measurement (SEM), and Standardized Indices of DIF (SID) for 29 Items

Change Type	*Item #*	*a*	SEM_a	b_{1993}	$SEM_{b\,(1993)}$	b_{1995}	$SEM_{b\,(1995)}$	*SID*
1.	1	1.120	.013	.531	.015	.797	.017	-11.733*
	2	1.188	.014	.686	.014	.932	.016	-11.571*
	3	1.115	.013	.759	.015	.774	.017	-.662
	4	.692	.008	.437	.022	.162	.026	8.074*
	5	.969	.011	.911	.017	.887	.020	.914
2.	6	.448	.005	.016	.034	.501	.040	-9.239*
	7	.975	.011	.657	.016	.637	.019	.805
	8	.949	.010	.034	.017	-.054	.020	3.353*
	9	.401	.005	-1.176	.039	-.563	.044	-10.426*
	10	.408	.005	-.452	.037	-.112	.043	-5.994*
	11	.846	.009	-.044	.019	-.027	.022	-.585
	12	.807	.009	.344	.019	.511	.023	-5.598*
	13	.882	.010	.827	.018	1.136	.021	-11.172*
	14	.839	.009	.283	.019	.589	.022	-10.527*
	15	.730	.008	-.402	.021	-.341	.025	-1.868
3.	16	.859	.009	.198	.018	-.184	.021	13.811*
	17	.742	.008	-.333	.021	-.671	.025	10.352*
	18	.691	.009	-1.586	.024	-1.721	.028	3.661*
	19	.903	.010	.758	.018	.669	.021	3.218*
	20	1.087	.012	.321	.015	.156	.017	7.278*
	21	1.126	.013	.849	.015	.380	.017	20.687*
	22	1.058	.012	.635	.015	.116	.018	22.150*
	23	1.125	.013	.617	.015	.326	.017	12.835*
	24	.896	.010	1.156	.018	.651	.021	18.258*
	25	1.066	.013	1.147	.016	.989	.018	6.561*
	26	1.095	.013	1.324	.015	.529	.017	35.066*
4.	27	.783	.009	-.897	.021	-.202	.023	-22.315*
	28	.751	.008	-.489	.021	-.078	.024	-12.888*
	29	.969	.011	.173	.017	.497	.020	-12.343*

a = item discrimination parameter, *b* = item difficulty/threshold parameter
* = SID, $p \leq .05$

DISCUSSION

The results of this study lend mixed support to the view that item anchors make a difference in the way survey items function and that their importance should not be overlooked when developing and using attitudinal measures. That is, significant DIF indices occurred for three of the five items that were exactly the same for both administrations of the survey. The hypothesis was that none of the five items would exhibit significant DIF between the 1993 and 1995 administrations. As item anchors can never be viewed independently from their item stems, likewise items must be considered in terms of the test or survey from which they are a part. Since the surveys across the two administrations differed somewhat, it is possible that this context difference strongly influenced the respondents' perceptions of the instrument. In addition, changes within the organization itself could serve to impart contextual effects on the administration of a survey. In particular, targeted organizational interventions were undertaken between the two survey administrations that may have not only impacted the construct being measured, but the items' ability to measure that which they were designed to measure. In short, it is possible that more than just the effects of the change in item anchors was impacting the stability of item parameters.

Before proceeding to a discussion of the implications, however, a few caveats and comments are necessary. The first of these concerns the use of IRT models in general. As stated by Hambleton and Swaminathan (1985), when there is a close fit between a chosen IRT model and the test dataset of interest, IRT models have a number of desirable qualities. These include: a) item parameter estimates that are independent of the particular sample of respondents that was used for calibration, b) respondents' latent trait estimates that are independent of the particular sample of items that was used for calibration, and c) a statistic indicating the precision of ability estimates. However, as with all mathematical models, IRT models include assumptions that are important for determining the adequacy of a particular model's fit to a dataset. Though the particular assumptions vary as a function of the type of model, some of the more common assumptions include: a) the modeled items purport to measure what constitutes a unidimensional construct, b) a respondent's answers to different items in a survey are statistically independent of each other (the assumption of local independence), c) the set of modeled items represent the complete latent space as it is defined operationally, and d) the survey is not speeded (Hambleton & Swaminathan, 1985).

The particular set of items that were modeled for this study was taken from a single survey, albeit one that was revised between administrations. However, the items in this study represent components of a number of different subsections that were not intended to be unidimensional when they were created. With this in mind, the dimensionality of the items was assessed. An exploratory principal components analysis revealed that the majority of these items loaded on a single component, indicating at least some degree of unidimensionality. Though a single model was used for this study, the decision to use such was at the discretion of the researchers. There is disagreement in the literature with regard to the robustness of particular IRT models to violations of unidimensionality and other IRT assumptions. Though Lord (1952) explicitly stated the primacy of the assumption of trait unidimensionality and some research (Ansley & Forsyth, 1985) has supported this assumption, other research (Drasgow & Hulin, 1990; Drasgow & Parsons, 1983; Hambleton, 1989; Harrison, 1986; Reckase, 1979) has shown that, with sufficient sample sizes and the dominance of one predominant dimension (in the cases where slight multidimensionality exists), IRT parameter estimates are stable and accurately represent the data. The degree to which this single model is actually a good representation of the data used in the present study must be interpreted with caution. The same is true with regard to the model's meeting the assumptions of local independence and specification of the complete latent space. The surveys upon which the data for this study are based were not administered under speeded conditions. Hence, the assumptions do not appear to be violated.

In addition to issues related to the assumptions of the IRT model used in this study, the conceptual bases of DIF, itself, need to be addressed. Though DIF can be viewed simply as a statistical difference between the item functioning characteristics of two groups of examinees to the same items, inherent in the conception of DIF is the idea of multidimensionality in the itemset. That is, in addition to the construct being measured by the survey items, at least one latent construct is influencing the item response

patterns between individuals. When a set of items is known to be essentially unidimensional, the detection of DIF is not generally as controversial as when the itemset is thought or is known to violate the assumption. For a unidimensional scale, DIF can be attributed to factors outside of what is being explicitly measured. When the scale dimensionality is in question, overall measurement quality is questionable.

A somewhat less urgent concern regarding DIF pertains to the way that the concept has been used in this study in relation to a more classical interpretation. When DIF was first conceived and when the initial procedures were developed, DIF analyses were associated mainly with differences in group performance that would have resulted from gender or racial-ethnic effects between groups of examinees. This was natural, given that DIF was originally considered an updated version of item bias. However, it is recognized that DIF pertains to statistical properties that may not be limited to those comparisons that originally drove bias studies. In other words, DIF is independent of the type of group comparison that is made. A good example of this is a specific type of DIF that is known as item drift (commonly represented as DRIFT in the literature). DRIFT exists in a test where the item parameters for a static test change over a period of successive administrations (Zimowski, Muraki, Mislevy, & Bock, 1996). Incidences of DRIFT are particularly common in large-scale testing situations (e.g., Graduate Record Examination) where information about the test may leak over a period of time. Despite this and other contrary examples, there is a belief among some (Tenopyr, 1994) that DIF is somehow tied to gender and racial-ethnic group membership. Given that the present study provides another alternative application of DIF methodology, concerns about appropriateness of the analyses may arise.

Another issue is whether the observed differences in item functioning are of practical significance. When properly defined and measured, DIF is assumed to be a result of variations in response generation at the cognitive level. Hence, DIF may be indicative of the presence of differential item/response category perception. Further, there are important implications for uniformity of scale scores and comparisons based on mean differences. Though for this dataset, statistically significant mean score differences existed for all but three items between the two administrations, the effect sizes for all but 11 of the 29 variables did not even attain a "small" level (Cohen, 1988), and only one could be considered a "medium" effect size. Statistically significant differences, or conversely, those that are not, and examination of score difference effect sizes are essentially meaningless without evidence that a particular measuring instrument is functioning with the same level of precision each time that it is used. However, the practical implication is determined by whether the modifications actually result in any change in the overall scale values and interpretation of the outcomes.

It is recognized that the post hoc measurement-based approach to the assessment of DIF used in this study may have been less than optimal. An alternative approach that is currently being investigated by the authors utilizes a controlled experimental design. Schmeiser (1982) has outlined procedures for conducting experimentally-based research for investigating the impact of item format on response patterns. Returning to the survey-based approach, another alternative involves the random assignment of alternate forms that would be distributed to comparable samples that had been sampled from the survey population of interest. This format would allow for statistical control of all conceivable study contaminants, and limit the variation to item-response anchor effects. Variations in item stems could be investigated in similar fashion. Further research should be conducted in the area of DIF (laboratory as well as field-based), with particular attention paid to an item's ability to discriminate between anchor categories and item category thresholds. However, pending further research, practitioners might best consider a conservative approach to item revision when item content remains the same.

REFERENCES

Alwin, D.F. (1997). Feeling thermometers versus 7-point scales. *Sociological Methods & Research, 25,* 318-40.

Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37-48.

Boudenhausen, G.V., & Wyer, R.S., Jr. (1987). Social cognition and social reality: Information acquisition and use in the laboratory and the real world. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 6-41). New York NY: Springer-Verlag.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks CA: Sage Publications.

Chang, L. (1997). Dependability of anchoring labels of Likert-type scales. *Educational and Psychological Measurement, 57,* 800-7.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale NJ: Erlbaum.

Dobson, K.S., & Mothersill, K.J. (1979). Equidistant categorical labels for construction of Likert-type scales. *Perceptual and Motor Skills, 49,* 575-80.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale NJ: Erlbaum.

Drasgow, F., & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol.2* (2nd ed., pp. 577-636). Palo Alto CA: Consulting Psychologists Press.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189-99.

DuBois, B., & Burns, J.A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement, 35,* 869-84.

Gable, R.K., & Wolf, M.B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings.* Boston MA: Kluwer-Nijhoff.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York NY: American Council on Education & Macmillan Publishing.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston MA: Kluwer-Nijhoff.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91-115.

Hofacker, C.F. (1984). Categorical judgement scaling with ordinal assumptions. *Multivariate Behavioral Research, 19,* 91-106.

Holland, P.W., & Wainer, H. (Eds.)(1993), *Differential item functioning.* Hillsdale NJ: Erlbaum.

Landy, F.J., Shankster, L.J., & Kohler, S.S. (1994). Personnel selection and placement. *Annual Review of Psychology, 45,* 261-96.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs, No. 7.* Chicago IL: University of Chicago.

Lynch, A., Barnes-Farrell, J.L., Kulikowich, J. (1998, April). *Do organizational survey items function differently for managers and non-managers?* Poster presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas TX.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-48.

Muraki, E. (1997, October 27). Personal communication.

Muraki, E., & Bock, R.D. (1997). *PARSCALE 3.0: IRT item analysis and test scoring for rating-scale data.* Chicago: Scientific Software, Inc.

Pace, C.R., & Friedlander, J. (1982). The meaning of response categories: How often is "Occasionally," "Often," and "Very Often"? *Research in Higher Education, 17,* 267-81.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-30.

Rockwood, T.H., Sangster, R.L., & Dillman, D.A. (1997). The effect of response categories on questionnaire answers. *Sociological Methods & Research, 26,* 118-40.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometric Monographs, 17.* Chicago IL: University of Chicago.

Schmeiser, C.B. (1982). Use of experimental design in statistical item bias studies. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 64-95). Baltimore MD: Johns Hopkins University Press.

Schriesheim, C., & Schreisheim, J. (1974). Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. *Educational and Psychological Measurement, 34,* 877-84.

Schwarz, N., & Hippler, H.-J. (1987). What response scales may tell your respondents: Informative functions of response alternatives. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 163-77). New York NY: Springer-Verlag.

Tenopyr, M.L. (1994). Big five, structural modeling, and item response theory. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.), *Biodata handbook* (pp. 519-34). Palo Alto CA: Consulting Psychologists Press.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory.* New York NY: Springer-Verlag.

Wright, B.D. (1979). *Best test design.* Chicago IL: University of Chicago Press.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago IL: Scientific Software, Inc.

APPENDIX A

Survey Items (29) Investigated in this Study (Classified by Item Change Type)

Type 1 – Identical Item Anchors

1. To what extent do you receive advance information from the FAA concerning major innovations or organizational changes that affect your job?

2. To what extent do you receive sufficient information from the FAA to understand how innovations and changes might affect you?

3. To what extent is your organization generally quick to use improved work methods?

4. To what extent have you had an opportunity to participate in FAA-funded training programs?

5. To what extent are there things about working in your organization (such as policies, practices, or conditions) that encourage you to work hard?

Type 2 – Similar Item Anchors (midpoint changed from "neither disagree nor agree" to "neutral")

1. I am required to get approval for decisions that I think I should be able to make myself.

2. Decisions in my organization are made at those levels where the most adequate and accurate information is available.

3. Management in my organization ensures that the information I need to do my job is readily available.

4. Some employees may be hesitant to speak up for fear of retaliation.

5. It is generally safer to say that you agree with management even when you don't really agree.

6. We are encouraged to express our concerns openly.

7. It's pretty common to hear "job-well-done" within my organization.

8. Promotions in my organization are given to those who are well qualified.

9. Rewards or recognition are given for exceptional performance in my organization.

10. I think THIS survey will provide top management with information on issues important to the workforce.

Type 3 – Response scales changed from "extent of..." to "agree/disagree"

1. I have been able to contribute to decision-making that affects my job.

2. I have the authority to make decisions to resolve most day-to-day work problems.

3. I understand how my job contributes to the FAA's mission.

4. Conflicts and differences in my organization are brought and managed, rather than avoided or worked around.

5. Policies and procedures affecting my work are communicated adequately.

6. The FAA is committed to people concerns.

7. My facility/organization has a real interest in the welfare and satisfaction of those who work here.

8. My facility/organization is effective in utilizing employees' skills and abilities.

9. Within the past 2 years, I have seen a positive change in the emphasis that the FAA places on managing people.

10. The FAA takes into account the impact of organizational changes on employees.

11. I believe that information from THIS SURVEY will be used by upper level management to improve working conditions and employee morale.

Type 4 – Response scales changed from "agree/disagree" to "extent of..."

1. To what extent have you been able to apply what you have learned from FAA training to your job?

2. To what extent have you received the training you need to perform effectively in your job?

3. To what extent does management in your organization use customer requirements and feedback to plan improvements in the products/services you provide?