*Edgeworth Series in Quantitative Educational and Social Science*
*(Report No.ESQESS-2001-3)*

# EFFECTS OF ITEM ORDER ON CONSISTENCY AND PRECISION UNDER DIFFERENT ORDERING SCHEMES IN ATTITUDINAL SCALES: A CASE OF PHYSICAL SELF-CONCEPT SCALES

**Charles O. Ochieng**

**University of British Columbia**

When citing this paper please use the following format (for APA style):

Ochieng, C. O. (2001). *Effects Of Item Order On Consistency And Precision Under Different Ordering Schemes In Attitudinal Scales: A Case Of Physical Self-Concept Scales* (Paper No. ESQESS-2001-3). Vancouver, B.C.: University of British Columbia. Edgeworth Laboratory for Quantitative Educational and Social Science.

Author note. A version of this paper was presented at the National Council for Measurement in Education conference in New Orleans, April 2000.

# EFFECTS OF ITEM ORDER ON CONSISTENCY AND PRECISION UNDER DIFFERENT ORDERING SCHEMES IN ATTITUDINAL SCALES: A CASE OF PHYSICAL SELF-CONCEPT SCALES

**Charles O. Ochieng**

**University of British Columbia**

## SUMMARY

In both classical and item response theory, item order has always been assumed to be invariant across test forms and instruments. However, cognitive theories (Schema theory) and information processing theories have shown that information is encoded and processed in a sequential and hierarchical pattern (Massaro & Cowan, 1993; Anderson, 1996). Therefore, responses are influenced by item order. In this regard the violation of the assumptions of item order invariance has implication to valid interpretation, precision, and development of instruments. While previous studies have addressed item order effects in achievement tests, little has been done in affective measures, and attitudinal scales in particular.

The purpose of this study is to examine the effect of item order on the consistency, precision and structure of attitudinal scales, and in particular, physical self-concept scale (PSCS). This study examines the effect of item order on score consistency, precision, and factor structure on the physical self-concept scale (PSCS) under three different ordering schemes. This resulted in three version of the instrument based on the proportion of endorsement within the categories of responses namely, ascending (AG), descending (DG) and random order (RG). The instrument was an adaptation of the PSCS scale (Reynolds, Flament, Masango, & Steele, 1999). The original instrument was validated with other instruments measuring self-concept (Rosenberg, 1965). Reliability coefficient of 0.86 was obtained for the scores in the validation study. The study sample consisted of 155 students, both male and female of ages 19 to 40years at the University of British Columbia.

While previous studies have addressed item order in achievement tests, little has been done in attitudinal scales. This study examines the effect of item order on Consistency, Precision and Factor structure of the PSCS under three different ordering schemes, with the ascending order scheme being the most consistent and the random order scheme being the least consistent. Precision and factor structure recovery varied more in random order scheme than in both ascending and descending order scheme. These findings have implications to valid interpretation of scores in attitudinal scales.

## Introduction

To measure psychological attributes, test items have been administered to respondents in various formats of item ordering. However, previous studies have shown that item order affect measurement and accurate interpretation of the instruments (Schur & Heriksen, 1983; Krosnick & Alwin, 1987; Chan, 1991; Zwick, 1991; Berger & Veerkamp, 1996; Sijtsma & Junker, 1996). This has implication to valid interpretations of the results obtained from the instruments.

The occurrence of item order effects can be attributed to item dependency, where responses to subsequent items are dependent or correlated to the previous items. In cognitive theory, item dependency can be accounted for by primacy and recency effects, as well as the development of schema. Thus, the link between cognitive process and measurement in psychometrics is illustrated through schema theory, which accounts for item order effects in responses.

### Psychometric concern

The main psychometric concern of item order effects is assumptions made in modeling responses in classical test theory (CTT) and item response theory (IRT). In CTT models, it is assumed that item response errors are uncorrelated and independent, while in IRT models, responses to items are assumed to be independent. This is termed as local independence in IRT models. The assumption of order invariance forms the basis for the application of these models in psychometrics.

In practical test situations and administration of instruments, violation of independence in IRT and the occurrence of correlated errors in CTT models are expected to occur as a result of cue utilization by respondents. Moreover, given the sequential and hierarchical nature of information processing (Anderson, 1981; Massaro & Cowan, 1993; Tatsuoka & Tatsuoka, 1996), responses to adjacent or subsequent items are expected to influence each other in both achievement and attitudinal scales. Based on this premises, it follows that item order will influence response patterns of individuals. This has implications to the validity and accuracy of interpretation of the results of instruments and hence the psychometric concern.

Current conventional test development practice does not incorporate item order effects in the test design, and little consideration is given to cognitive theory of learning and information processing. Therefore, there is lack of fidelity between cognitive constructs being measured and conventional test development practices (Nichols, & Sugrue, 1999). A lot of emphasis has been laid on mathematical models to account for responses at the expense of cognitive theories and information processing approaches that take item order into account. For valid interpretation of the responses and inferences made on the constructs being measured, both theoretical and psychological considerations on response patterns need to be considered. Given the sequential nature of information processing, association between the responses of adjacent items is bound to occur resulting in item dependency. This contradicts the assumption of the independence of the responses and poses a threat to valid inference of results. Therefore, item dependency has been identified as a psychometric concern given its implication to valid interpretations of the scores.

In recognizing the inevitable occurrence of item dependency, Yen (1984; 1993) recommended the use of testlets as a method of managing item dependency. Testlets, which are formed from the item bundles, result in shorter instruments. However, this method results in loss of information, besides the fact that reliability of the resulting shorter instrument is compromised. Statistical distributions of the indices proposed by Yen are also not known. Other methods of managing item dependency have also been proposed. Ackerman and Spray (1986) proposed models based on conditional probability that incorporated item dependency in IRT models, in accounting for response data. Ackerman (1987) later presented a model for response data based on Marchov Chains. Due to mathematical complexity of the models, the proposed method has had limited application. Thus, the issue of item dependency and order effects has yet to be resolved. While several studies have addressed the effect of item order in achievement tests (Leary & Doran, 1985, Kingston & Doran, 1984; Zwick, 1991; Sijtsma

& Junker, 1996) further investigation has yet to be done in attitudinal scales in this regard. Therefore, this study addresses item order effects in attitudinal scales with special reference to self-concept scales.

## Objective of the study

The objective of this study is to assess the effects of item order in attitudinal scales using a physical self-concept scale, the PSCS (Physical Self-Concept Scale) instrument (Reynolds, Flament, Masango, & Steele, 1999). The study investigated the effects of item order on the consistency of responses and accuracy of the scale under different item order schemes. Psychometric concerns have been raised so far regarding item order and its validity implications to the interpretation of data. This was addressed in the study. These findings have an impact on the development, validation and improvement of self-reporting scales.

## Theoretical Framework

Item order effects are studied by examining the change in responses with the change in item position or order of options available to the respondents. Reversing or changing the position of items and response options were found in previous studies to result in a change in response patterns, and in information processing. These changes can be accounted for by proximity and primacy effects (Waugh & Norman, 1965; Milligan, 1979; Kronick & Alwin, 1987; Solso, 1991) as well as information processing strategies adopted by respondents (Massaro & Cowan, 1993). Primacy effects occur when placing an item at the beginning of a list increases the likelihood of the item being respondent to favorably. An explanation of this occurrence is that the earlier items anchor a cognitive framework and serves as a basis upon which information processing required to respond to the subsequent items are based.

Hierarchical and sequential nature of information processing influence the response patterns for a given set of items. It follows that items adjacent to each other are expected to have responses that are highly correlated due to proximity. This is termed as proximity effect, which often confound with primacy effect for adjacent items. Therefore, in studying item order effects through primacy and proximity effects, the pattern of the inter-item correlation matrices are tested for invariance across different ordering schemes. First, second and third lag effects are also examined to determine the effect of item order through primacy and proximity effects, on the correlation matrices and the test structure of the instrument.

Krosnick and Alwin (1987) developed a theoretical model to explain the underlying processes of responding to scales. According to this model, individuals are expected to respond to the first satisfactory option in order to minimize cognitive effort rather than perform an exhaustive search for optimal solution. Responses to subsequent items are based on previous experiences and responses to previous items. Krosnick and Alwin (1987) found that item responses of subsequent items were correlated, and dependent on each other. They observed that the closer the items in terms of the spatial distance, the higher the correlation between the respective responses.

Kronick and Alwin (1987) tested the hypothesis that inter-item correlation was invariant across item order forms. They found that both variances and covariance of the items varied with different item order forms. This result was explained in terms of primacy and proximity effects. In proximity effects, responses to adjacent items tend to have a higher correlation among them than they have with responses to other items. Therefore, inter-item correlation is hypothesized to depend on the ordering of items due to proximity and primacy effects. This is expected to have an affect on the factor structure of the instrument.

## Purpose of the study

The purpose of the study is to investigate the effects of item order on the consistency of responses, factor structure, and accuracy of attitudinal scales under different item ordering schemes using the PSCS scale of self-concept (Reynolds et.al 1999). Findings are generalized to self-concept, and attitudinal scales. Three item-ordering schemes are studied to determine proximity and primacy effects, which are a manifestation of item order in responses. Based on the current psychometric and research concerns, the study answers three research questions.

## Research Questions

1.What are the effects of item order on the consistency of responses across different ordering schemes?

2. What are the effects of item order on the variance-covariance structure of the responses across different ordering schemes?

3. What are the effects of item order on the relative precision of the instrument across different ordering schemes?

## Method

Participants were recruited from the undergraduate and graduate student population at the University of British Columbia. The sample consisted of both male and female students from different cultural background and ages ranging from 19 to 40 years. Three versions of the instrument representing the three ordering schemes were randomly assigned to participants to create a randomized three-group study design.

### Instrument

An adaptation of the physical self-concept scale PSCS (Reynolds et.at 1999) was administered to participants. This particular instrument was selected because of its sound psychometric properties in terms of reliability and validity evidence. To determine evidence of validity, the original instrument was cross-validated with other self-concept scales and found to be a valid measure of physical self-concept. When correlated to the Rosenberg Self-esteem scale, RSES (Rosenberg, 1965) a correlation of 0.8 resulted, indicating a high level of criterion related evidence of validity. The instrument was found to have a reliability coefficient of 0.86. Self-concept construct was selected for study, as it is relatively stable and well documented (Shavelson, Hubner, & Stanton, 1976). This would enhance generalization of the findings to other attitudinal measures.

### Sample size

Previous studies on item order effect have not directly reported effect sizes. However, given the statistics used to compare item order forms, and the level of statistical significance in the studies, it can be generalized that low to moderate effect sizes are common in studies on item order effects. Table 1 shows effect sizes obtained from a sample of previous studies on item order. It is observed that effect sizes across the studies range from moderate to low (Cohen & Cohen, 1988).

Table 1
Effect sizes from selected studies on item order effects

| Study | Item order comparison | Statistics and level of significance ($p<0.01$) | Cohen's f (effect size) | Eta squared $\eta^2$ |
|---|---|---|---|---|
| Schurr & Henriksen (1983) | Test for proximity and order effects for three ordering schemes. Lag effects are tested for statistical significance. | $\chi^2$ (91) =198.86 | 0.440 | 0.166 |
| | | $\chi^2$ (153) = 327.58 | 0.572 | 0.247 |
| | | $\chi^2$ (171) = 471.36 | 0.680 | 0.321 |
| Kronick & Alwin (1987) | Test for form or item order invariance for two samples and two test order forms | $\chi^2$ (12) =31.86 | 0.178 | 0.031 |
| | | $\chi^2$ (66) = 107.94 | 0.320 | 0.097 |
| Chan (1991) | Fitting one-factor and two-factor models for an ordering scheme and a reversed form of the scheme. | $\chi^2$ (5) = 24.01 | 0.155 | 0.023 |
| Burns (1996) | Tested item order effects on the consistency of responses for two groups. | F (1,34) = 13.54 | 0.630 | 0.285 |

Pearson's product moment correlation from which effect sizes can be derived was used to compare the three ordering schemes in this study. A moderate inter-item correlation ranging from 0.3 to 0.6 is expected to occur among the items based on the trends of the previous studies. Therefore, a moderate correlation of about 0.4 with an associated power of 0.8 would require a sample size of approximately 50 at an alpha level of 0.05 (Kreamer & Thiemann, 1987). This served as the rationale for selecting the sample size in each ordering scheme. To control for attrition, participants were over-sampled to 75 per group of item ordering scheme. Given the statistics used to compare item order effects in previous studies, the average effect size implied by correlation values ranged from low to moderate. For this reason, it was expected that a moderate effect size would occur in this study.

For ascending order group AG, the return rate was 42 out of 75. In the case of descending order group DG, the return rate was 58 out of 75, and for random order group RG, 55 out of 75. The total sample size was 155. Each group was then analyzed separately to determine the effect of item order on the three variables namely, consistency, precision, and factor structure of the scale under the stated conditions.

Criterion for ordering the items

Ordering of items in the adapted instrument was based on the endorsement data from the validation study of the original instrument measuring physical self-concept (Reynolds et.al 1999). The proportion of "agree" endorsement was adopted as a criterion of ordering because it is analogous to the concept of item difficulty in classical test theory. Moreover, a high proportion of endorsement implies that the item has a high correlation with the construct or attribute being measured. This also corresponds to high threshold values.

Given the occurrence of primacy effects in the ordering schemes, inter-item correlation is expected to be significantly different between the ordering schemes based on the criterion of high to low proportion of "agree" endorsement, and from low to high proportion of "agree" endorsement. The ordering criterion is based on the endorsement information from the validation data of the original instrument (see Table 2).

The proportion of "agree" endorsement implies a high correlation between the items and the attribute the instrument is measuring. Inter-item correlation among items and internal consistency are hypothesized to be statistically, significantly different among item ordering schemes where items are systematically, and randomly ordered for low and high proportion of "agree" endorsement.

Table 2
Item statistics indicating proportion of endorsement of item agreement, mean and standard deviation, extracted from the validation study data

| Items | Direction of wording | Domain | %Proportion Endorsed "Agree" | Item Mean | Standard deviation |
|---|---|---|---|---|---|
| Item 1 | + | PA | 69 | 3.08 | .55 |
| Item 2 | - | PA | 48 | 2.83 | .78 |
| Item 3 | + | PA | 57 | 2.66 | .70 |
| Item 4 | + | PA | 69 | 2.84 | .57 |
| Item 5 | + | PA | 54 | 2.66 | .68 |
| Item 6 | + | PA | 40 | 2.64 | .80 |
| Item 7 | + | PA | 73 | 2.89 | .54 |
| Item 8 | + | PA | 47 | 2.66 | .76 |
| Item 9 | + | PA | 54 | 2.91 | .76 |
| Item 10 | + | PA | 26 | 2.25 | .77 |
| Item 11 | + | PS | 60 | 2.95 | .67 |
| Item 12 | + | PS | 57 | 2.74 | .68 |
| Item 13 | - | PS | 50 | 3.09 | .74 |
| Item 14 | + | PS | 57 | 3.10 | .70 |
| Item 15 | + | PS | 48 | 2.88 | .78 |
| Item 16 | + | PS | 51 | 3.13 | .56 |
| Item 17 | + | PS | 69 | 3.11 | .56 |
| Item 18 | - | PS | 45 | 2.78 | .76 |
| Item 19 | + | PS | 58 | 2.76 | .67 |
| Item 20 | - | PS | 44 | 3.11 | .78 |

Sample size, n= 654. Domains selected from the instrument are PA (Physical Appearance), and PS (Physical Skill/ability).

Item ordering schemes in the instrument

Twenty items from the original instrument (PSCS) were selected and ordered in three ordering schemes. Ten items were selected from physical appearance (PA) domain and ten from physical ability (PS) domain in PSCS. This resulted in three versions of the original instrument with twenty items each. The first ordering scheme, referred to as DG, was ordered from high proportion of "agree" endorsement to low proportion of "agree" endorsement. Each domain was ordered separately to avoid possible contamination and confounding of group effects based on the two domains.

In the second ordering scheme, items were ordered from low proportion of "agree" endorsements to high proportion of "agree" endorsement. This is referred to as AG. In the third ordering scheme, items with high and low proportion of "agree" endorsement were randomly ordered (using table of random numbers) and the resulting version referred to as RG. This version was used as a baseline to which responses from the other two ordering schemes were compared. Table 3 shows a schematic presentation of the three ordering schemes. To test the hypotheses of overall effects of item order, consistency of responses and the equivalence of the pattern of the Inter-item correlation matrices, were determined and compared among the three ordering schemes according the planned contrast, based on the theory and previous studies.

Table 3

Schematic presentation of criterion of the three item ordering schemes

| Instrument versions | DG<br>Descending order scheme. | AG<br>Ascending order scheme. | RG<br>Random order scheme. |
|---|---|---|---|
| Ordering schemes | Ungrouped items arranged from high to low proportion of "agree" endorsement of items. | Ungrouped items arranged from low to high proportion of "agree" endorsement of items. | Ungrouped items randomly arranged with low, moderate and high proportions of "agree" endorsement of items. |

**Hypotheses**

It was hypothesized that responses to the items vary with different item ordering schemes, and that variability in responses result in variability in the measures of consistency and information across the ordering schemes. Factor structure recovery was also hypothesized to vary across the three different ordering schemes. Differences in test structure were determined by testing for equivalence of the inter-item correlation matrices across the three ordering schemes. Expected variability in responses in the ordering schemes can be accounted for by the occurrence of primacy and proximity effects.

First lag effect in the inter-item correlation matrices indicated the presence and magnitude of primacy effects as a result of item ordering. This was tested across the three inter-item correlation matrices. The second and third lag effects tested across the matrices indicated the magnitude and significant differences of proximity effects across the ordering scheme. The overall order effect on the test structure was determined by testing for the equivalence of the pattern of the inter-item correlation matrices. This measured the lag effect for all the items in the correlation matrices.

It was hypothesized that proximity and primacy effects are lower in the random ordering scheme RG, than in the two other schemes namely, AG and DG, because in the random scheme, it is assumed that the items are independent as a result of randomization. In AG and DG, the systematic ordering schemes were expected to result in items influencing subsequent items. Items that are adjacent are expected to correlate highly than items positioned far apart in the ordering scheme. This corresponds to spatial distance among the items. Therefore, inter-item correlations in the first lag (elements in the first diagonal, immediately below the main diagonal in the inter-item correlation matrix) in AG and DG were expected to be greater than those in RG.

In the case of items positioned far apart, the correlation can be largely accounted for by primacy effects. Second and third lag inter-item correlation in AG and DG were expected to be larger than those in the random ordering scheme RG. A similar trend was expected for the overall lag effects across the inter-item correlation matrices. This was determined by testing for the equivalence of the pattern of correlation matrices across the three groups. No statistically significant proximity effects were expected in the random ordering scheme RG. A schematic presentation of the comparison of the first, second, and third effects in the inter-item correlation across the three item ordering schemes is shown in Table 4. Item ordering scheme RG was used as a baseline to which each of the other ordering schemes were compared. From the first lag effects, proximity effects were tested for significance across the ordering schemes. The following hypotheses were tested on item order effects through inter-item correlation in the ordering schemes.

**Hypothesis 1.1:** Proximity effect indicated through the first lag inter-item correlation among items in the first ordering scheme AG is significantly greater than in the random ordering scheme RG. The hypothesis was tested by comparing the mean of the first lag inter-item correlation $r_{11}$ in AG to $r_{13}$ in RG.

**Hypothesis 1.2:** Proximity effect indicated through the first lag inter-item correlation in the ordering scheme DG, is significantly greater than in the random ordering scheme RG. This hypothesis was tested by the comparing the mean of the first lag inter- item correlation $r_{12}$ in DG to $r_{13}$ in RG.

The three ordering schemes were tested for primacy effects using second and third lag inter-item correlation. Second lag inter-item correlation across the three ordering schemes were expected to be lower than first lag inter-item correlation because of the diminishing influence of the subsequent items, due to lower proximity. Thus, the only influence expected is primacy effect. However, second lag inter-item correlations were expected to be larger than third lag inter-item correlations for the same reason. Hypotheses of primacy effects across the ordering schemes are presented below. For the second lag inter-item correlation across the AG and RG ordering schemes, the substantive hypothesis is stated as follows.

**Hypothesis 1.3:** The primacy effects in the first ordering scheme AG is significantly greater than in the random ordering scheme RG. This hypothesis was tested by comparing the mean of the second lag inter-item correlation $r_{21}$ in AG to $r_{23}$ in RG.

**Hypothesis 1.4:** The primacy effects in the second ordering scheme DG is significantly greater than those in the random ordering scheme RG. The hypothesis was tested by comparing the mean second lag inter-item correlation $r_{22}$ in DG to $r_{23}$ in RG. For the third lag effects, the mean of the third lag inter-item correlation were compared between AG and RG, and DG and RG. However, the third lag effects indicated by the inter-item correlation is expected to be lower than in the case of second lag effects. Hypotheses on primacy effects based on the third lag effects are presented as follows.

**Hypothesis 1.5:** Primacy effect in the ordering scheme AG is significantly greater than that in the random ordering scheme RG. This hypothesis was tested by comparing the mean of the third lag inter-item correlation $r_{31}$ in AG to $r_{33}$ in RG.

**Hypothesis 1.6:** Primacy effects in the second ordering scheme DG is significantly greater than that in random order scheme RG. The hypothesis is tested by comparing the mean of the third lag inter-item correlation $r_{32}$ in DG to $r_{33}$ in RG. In order to control for an inflated type one error rate across the six hypotheses an overall alpha was set at $p<0.05$ according to Bonferroni procedure. Table 4 shows the comparison of the means of the inter-item correlation and the lag effects across the ordering schemes.

Table 4

Comparison of the means of the inter-item correlation and lag effects across ordering schemes

| Inter-item correlation | Lag effect | AG | DG | RG | Item order effect |
|---|---|---|---|---|---|
| First lag correlation | 1 | $r_{11}$ | $r_{12}$ | $r_{13}$ | High Proximity and primacy |
| Second lag correlation | 2 | $r_{21}$ | $r_{22}$ | $r_{23}$ | Moderate proximity and primacy |
| Third lag correlation | 3 | $r_{31}$ | $r_{32}$ | $r_{33}$ | Low proximity and primacy |

In each ordering scheme, the reliability of the scores measured by each of the instrument version was computed. The resulting alpha coefficient of consistency across the three ordering schemes were tested for significant differences to determine the effect of item order on the consistency of the instrument versions.

Standard errors of the mean of each item, in each ordering scheme were also computed and the test information function obtained for each ordering scheme. Relative precision of AG and DG was computed by comparing the information function of AG to RG, and DG to RG.

## Data Analysis

Descriptive statistics was computed for each group of ordering scheme. Standard errors of the item means were also computed to infer statistical information at each item level as well as score level.

To satisfy the statistical assumption required for factor analysis, a p-p plot was conducted to determine if there was violation of normality assumption. For violation of multivariate normality, Bartlett's test of sphericity was conducted. Kaiser-Meyer-Olkin test for sampling adequacy was also conducted. In both cases the data was found to be suitable for factor analysis.

To test for similarity of correlation and covariance structure across the three groups' matrices, Kaiser's method of average correlation for corresponding sub-diagonals was conducted by transforming r-values to Fisher's z. For comparison of consistency, alpha coefficients across the groups were also tested for statistically significant differences at 0.01 level of significance (Charter, 1997; Feldt, Woodruff, & Salih, 1987). To compare precision across the ordering schemes, relative precision was computed from the ratio of the test statistical information for each version of the instruments AG and DG, to that of random order scheme RG.
The following formulae derived from Feldt, Woodruff and Salih (1987) are used to test for the differences for two or more independent reliability coefficients' alpha and Kuder-Richardson's (KR-20). The sample coefficient alpha, including the KR-20, is denoted by A. The formulas for testing for the significant differences between *k* independent alphas with *n* number of test items and *N* number of subjects for *i* tests are

$$B_i=N_i\,(n_i-1)/n_i+1) \qquad (1)$$

$$C_i = 2/[9(B_i-1)(1-A_i)^{2/3}] \qquad (2)$$

$$D= S\,C_i/k \text{ (i.e the mean of the } C_i \text{ s)} \qquad (3)$$

$$E = S[1/(1-A_i)^{1/3}]/k \qquad (4)$$

$$c^2 = S[1/(1-A_i)^{1/3}\ E]^2/D \qquad (5)$$

The degrees of freedom (*df*) for the chi-squared value is *k-1*.
If the overall chi-square is significant, the pairwise post hoc comparisons ($H_0: A_i = A_j$) are made by the

following F tests described in Charter and Feldt, (1996):

$$F_{ij} = (1-A_i)/(1-A_j) \qquad (6)$$

Where $A_i < A_j$ with degrees of freedom $N_j-1$ and $N_i-1$. The post hoc uses the Bonferroni adjustment to control for type one experimentwise error rate at $\alpha_e$. This adjustment is made by dividing $\alpha_e$ by *j* number of comparisons, where $j = (k^2 - k)/2$ for *k* number of alphas. The formulas (1) to (6) were originally developed by Feldt , Woodruff and Salih, (1987).
In this study, the average correlation in the inter-item correlation matrix was calculated using Kaiser's method, based on the largest eigenvalue and the number of variables in the correlation matrix. In Kaiser' s method, the average correlation $\gamma$ is obtained by dividing the largest eigenvalue $\lambda$ minus one, by the number of variables ***r*** minus one, as shown below.

$$\gamma = \frac{l-1}{r-1}$$

The average correlation obtained is then transformed into a Fisher’s z value. Similarly, the average correlation of the other matrices are obtained and also transformed into Fisher’s z values. The standard error of the population Fisher’s z is then computed. A test of significance differences between the two Fisher’s z values in conducted, based on the mean of the two Fisher's z values and the standard error of the two samples. Standard error of Fisher's z is computed from the following expression;

$$s_{zd} = \sqrt{\left(\frac{1}{N_i - 3}\right) + \left(\frac{1}{N_j - 3}\right)}$$

Where $s_{zd}$ is the standard error of the Fisher's z values for the samples *i* and *j* of which the differences are being tested for statistical significance. $N_i$ and $N_j$ denote the sample sizes of samples *i* and *j* respectively. Evidently, the standard error is dependent on the sample sizes being tested.

The computation of the information function in the three ordering schemes was derived from the standard errors based on the expression of the relationship of Shannon information index, $I\ (q)$ and the standard error of measurement, SEM shown as follows:

$$\text{SEM} = \frac{1}{\sqrt{I\,(q\,)}}$$

and

$$I(q) = \frac{1}{(SEM)2}$$

## Results and discussion

Comparison of the mean, standard deviation, consistency, and precision are summarized in Table 2. The mean and standard deviation of AG and DG were larger than those of RG. Using Charters method (Charter, 1997), differences in reliability coefficients were found to be significant ($\chi^2_{(2)}$ =12.18, p<0.01). F-test post hoc results indicated significant differences between AG and RG, and, DG and RG.

There were no statistically significant differences in the pattern of correlation matrices across AG, DG and RG. However, as Table 5 and figure 1 show, there were changes in the trends across the lags. Failure to detect differences is due to low sample size upon which the standard error of Fisher’s z statistic depends. Significant differences may be detected with large sample size than with low sample sizes as the standard error of Fisher's z is a direct function of the square root of the sample size. For each lag, the mean of z decreased across the groups from AG to RG (see Table 6).

Factor analysis (principal axis factor analysis) indicated different factor structure recovery across AG, DG and RG (see Table 7). Using the rule of extracting factors with eigenvalues greater than one, resulted in four factors in AG, and six factors in both DG and RG. However, using the Scree-plot method, two factors were extracted in all the three groups.

Table 5

Comparison of distribution, consistency and precision across the three groups

| Order scheme | Mean M | *SD* | Reliability Coefficient alpha | SEM | Test information | Relative precision |
|---|---|---|---|---|---|---|
| AG | 56.60 | 9.59 | 0.84 | 3.84 | 0.068 | 1.42 |
| DG | 56.14 | 9.09 | 0.79 | 4.17 | 0.058 | 1.21 |
| RG | 55.63 | 6.63 | 0.53 | 4.55 | 0.048 | 1.00 |

Table 6

Comparison of lag effects, mean correlation r, and Fisher's z in the correlation matrices across the three groups

| Lag effects | AG (ascending) | | DG (descending) | | RG (random) | |
|---|---|---|---|---|---|---|
| Lag 1 | $r_{11}$=.51 | $z_{11}$=.56 | $r_{21}$=.50 | $z_{21}$=.55 | $r_{31}$=.41 | $z_{31}$=.44 |
| Lag 2 | $r_{12}$=.48 | $z_{12}$=.53 | $r_{22}$=.44 | $z_{22}$=.48 | $r_{32}$=.39 | $z_{32}$=.42 |
| Lag 3 | $r_{13}$=.51 | $z_{13}$=.56 | $r_{23}$=.37 | $z_{23}$=.38 | $r_{33}$=.38 | $z_{33}$=.40 |

Table 7

Comparison of the factor structure recovery in the three groups, AG, DG, and RG

| Ordering scheme | | Factor Extraction Method | | |
|---|---|---|---|---|
| Group | Scree-plot Method | %Variance explained | Eigenvalue>1 Method | %variance explained |
| AG | 2 factors | 49.4 | 4 factors | 65.34 |
| DG | 2 factors | 49.0 | 6 factors | 66.30 |
| RG | 2 factors | 44.8 | 6 factors | 61.20 |

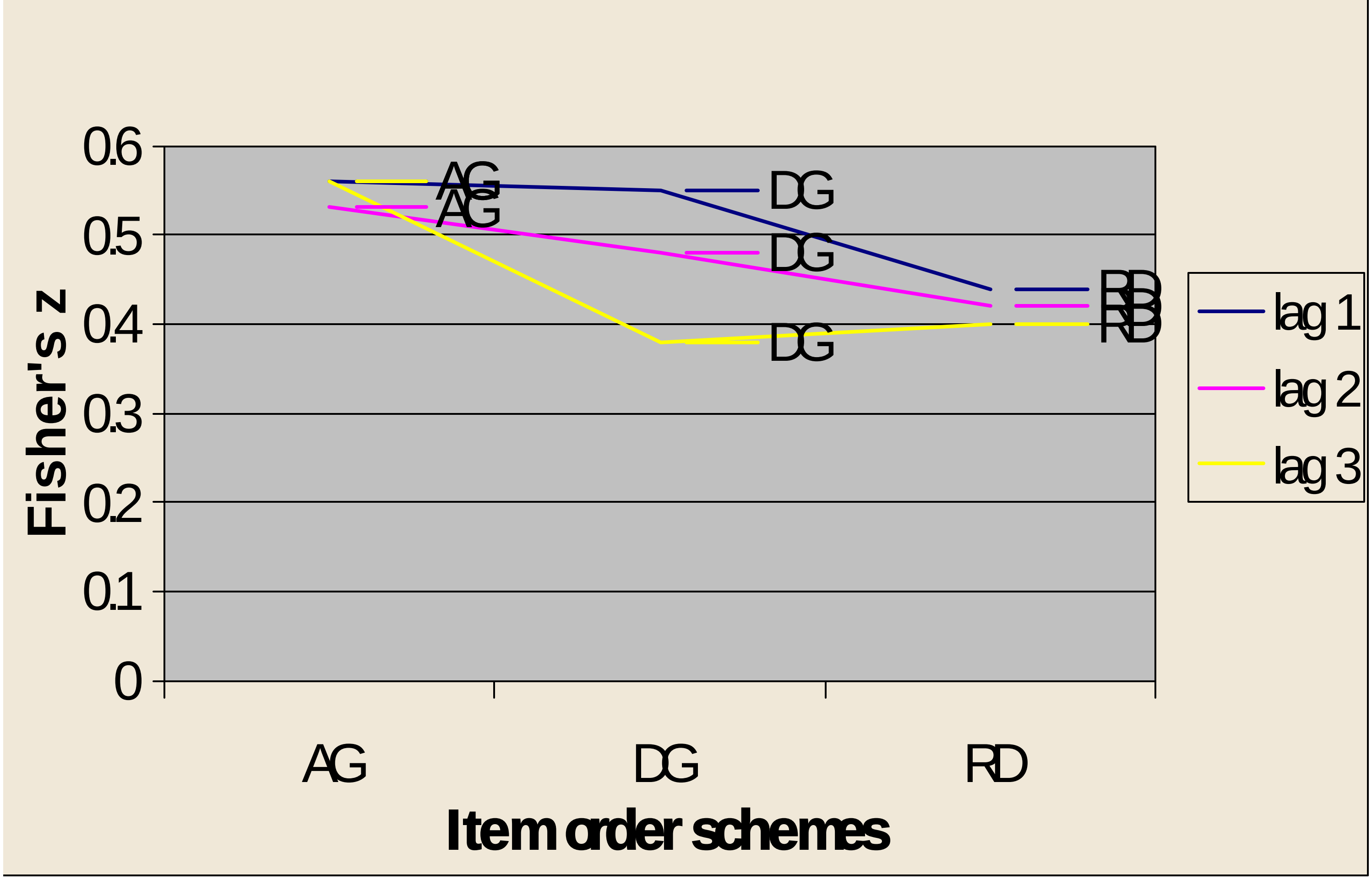
## Conclusion

Consistency varied with item order and was lowest in random ordering scheme, and highest with ascending scheme. This was evident in the size of the alpha coefficients in AG, DG and RG, which systematically decreased in that order (insert alpha values). As evidenced by the line graph the pattern of the consistency of the alpha values indicated a downward trend from AG to RG

Relative precision was highest in ascending scheme, and lowest in random scheme. A similarly decreasing trend was observed for relative precision from AG to RG. The RG ordering scheme has the lowest precision and consistency among the three ordering schemes, and yet it is a common psychometric practice to randomize items in an instrument. These results have implication to the present practice of randomizing items in instrument design and scale development. It is recommended that sequenced information processing which influence response pattern being incorporated in instrument development and design, for a meaningful and accurate interpretation.

Factor structure recovery was better in ascending order scheme than in the other two schemes.

Fig.1. Trends of lag effects across the three groups



There is a direct relationship between sequential information processing, response consistency, and precision of attitudinal scales. This needs to be considered in test design, development and interpretation, in order to incorporate cognitive theories to psychometric practice.

Table 8

Inter-item correlation matrix for Descending ordering scheme DG

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | | | | | | | | | | | |
| 2 | 648 | 1 | | | | | | | | | | | | | | | | | | |
| 3 | 753 | 752 | 1 | | | | | | | | | | | | | | | | | |
| 4 | 284 | 169 | 144 | 1 | | | | | | | | | | | | | | | | |
| 5 | 458 | 537 | 461 | 220 | 1 | | | | | | | | | | | | | | | |
| 6 | 194 | 325 | 311 | 202 | 343 | 1 | | | | | | | | | | | | | | |
| 7 | 217 | 468 | 409 | 001 | 536 | 288 | 1 | | | | | | | | | | | | | |
| 8 | 801 | 596 | 729 | 191 | 485 | 157 | 328 | 1 | | | | | | | | | | | | |
| 9 | 685 | 712 | 819 | 198 | 461 | 334 | 508 | 708 | 1 | | | | | | | | | | | |
| 10 | 607 | 679 | 645 | 257 | 490 | 285 | 410 | 549 | 669 | 1 | | | | | | | | | | |
| 11 | 368 | 289 | 448 | 143 | 214 | 115 | 207 | 424 | 403 | 475 | 1 | | | | | | | | | |
| 12 | 197 | 208 | 185 | 099 | 054 | 239 | 228 | 219 | 217 | 144 | 400 | 1 | | | | | | | | |
| 13 | 384 | 388 | 451 | 056 | 329 | 229 | 357 | 505 | 417 | 269 | 215 | 517 | 1 | | | | | | | |
| 14 | -03 | 417 | 038 | -19 | 078 | 193 | 006 | 046 | 037 | -74 | -15 | 087 | 083 | 1 | | | | | | |
| 15 | 492 | 553 | 562 | 172 | 511 | 235 | 587 | 562 | 543 | 600 | 516 | 378 | 436 | 138 | 1 | | | | | |
| 16 | 434 | 408 | 444 | 091 | 485 | 041 | 452 | 486 | 362 | 493 | 375 | 243 | 296 | 195 | 748 | 1 | | | | |
| 17 | 478 | 465 | 504 | 189 | 517 | 223 | 498 | 537 | 536 | 505 | 443 | 468 | 549 | 206 | 823 | 748 | 1 | | | |
| 18 | 329 | 343 | 372 | 171 | 338 | 175 | 346 | 394 | 389 | 295 | 277 | 484 | 359 | 184 | 486 | 530 | 607 | 1 | | |
| 19 | 447 | 439 | 446 | 121 | 393 | 209 | 379 | 449 | 403 | 524 | 393 | 305 | 370 | 009 | 711 | 637 | 737 | 480 | 1 | |
| 20 | 238 | 435 | 510 | -09 | 353 | 204 | 546 | 364 | 466 | 471 | 534 | 327 | 337 | 221 | 680 | 522 | 599 | 478 | 596 | 1 |

Table 9
Inter- item correlation matrix for random ordering scheme, RG

| items | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | | | | | | | | | | | |
| 2 | 626 | 1 | | | | | | | | | | | | | | | | | | |
| 3 | 410 | 249 | 1 | | | | | | | | | | | | | | | | | |
| 4 | 381 | 535 | 200 | 1 | | | | | | | | | | | | | | | | |
| 5 | 039 | 138 | 383 | 279 | 1 | | | | | | | | | | | | | | | |
| 6 | -099 | -074 | 022 | 009 | 564 | 1 | | | | | | | | | | | | | | |
| 7 | 504 | 480 | 458 | 513 | 344 | 076 | 1 | | | | | | | | | | | | | |
| 8 | 313 | -214 | 013 | -165 | 490 | 449 | -312 | 1 | | | | | | | | | | | | |
| 9 | 342 | 316 | -006 | 179 | -091 | -053 | 104 | 092 | 1 | | | | | | | | | | | |
| 10 | 126 | 187 | 244 | 261 | 599 | 543 | 312 | 363 | 031 | 1 | | | | | | | | | | |
| 11 | 031 | -048 | 200 | -076 | 027 | 078 | 021 | 298 | 032 | 206 | 1 | | | | | | | | | |
| 12 | 122 | 072 | 262 | -078 | 147 | -008 | 021 | 373 | 113 | 148 | 496 | 1 | | | | | | | | |
| 13 | 048 | 103 | 244 | -121 | 287 | 333 | 060 | 518 | 343 | 372 | 366 | 578 | 1 | | | | | | | |
| 14 | 152 | 273 | 433 | 521 | 185 | -098 | 109 | 310 | 083 | 333 | 433 | 618 | 551 | 1 | | | | | | |
| 15 | 095 | 212 | 063 | 037 | 120 | -123 | 026 | 150 | 178 | 228 | 257 | 587 | 357 | 432 | 1 | | | | | |
| 16 | -088 | -042 | 133 | -133 | 404 | 378 | 006 | 558 | 068 | 445 | 556 | 600 | 679 | 490 | 358 | 1 | | | | |
| 17 | -138 | -068 | 264 | -126 | 218 | 182 | 006 | 327 | 007 | 095 | 448 | 732 | 450 | 340 | 432 | 496 | 1 | | | |
| 18 | -060 | 123 | 243 | 031 | 240 | 171 | 018 | 456 | 143 | 272 | 517 | 757 | 665 | 496 | 572 | 693 | 778 | 1 | | |
| 19 | 020 | -124 | -247 | -102 | -255 | -174 | -186 | -104 | 082 | -221 | -346 | -486 | -360 | -280 | -300 | -478 | -566 | -66 | 1 | |
| 20 | 168 | 047 | 335 | -076 | 247 | 301 | 138 | 410 | 231 | 286 | 488 | 488 | 584 | 307 | 241 | 679 | 449 | .596 | -.39 | 1 |

Table 10

Inter-item correlation matrix for the ascending order scheme AG.

| item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | | | | | | | | | | | |
| 2 | 614 | 1 | | | | | | | | | | | | | | | | | | |
| 3 | 434 | 449 | 1 | | | | | | | | | | | | | | | | | |
| 4 | 716 | 487 | 484 | 1 | | | | | | | | | | | | | | | | |
| 5 | 583 | 302 | 418 | 481 | 1 | | | | | | | | | | | | | | | |
| 6 | 764 | 490 | 448 | 629 | 541 | 1 | | | | | | | | | | | | | | |
| 7 | 204 | 276 | -28 | 159 | 184 | 194 | 1 | | | | | | | | | | | | | |
| 8 | 654 | 593 | 308 | 770 | 498 | 515 | 301 | 1 | | | | | | | | | | | | |
| 9 | 914 | 601 | 441 | 764 | 629 | 775 | 173 | -69 | 1 | | | | | | | | | | | |
| 10 | 276 | 376 | 059 | 556 | 286 | 185 | 450 | 612 | 352 | 1 | | | | | | | | | | |
| 11 | 600 | 392 | 203 | 572 | 315 | 496 | 199 | 445 | 559 | 307 | 11 | | | | | | | | | |
| 12 | 651 | 645 | 320 | 671 | 356 | 654 | 382 | 640 | 694 | 453 | 711 | 1 | | | | | | | | |
| 13 | 544 | 444 | 021 | 584 | 283 | 486 | 309 | 492 | 625 | 400 | 554 | 574 | 1 | | | | | | | |
| 14 | 518 | 406 | 155 | 549 | 363 | 364 | 207 | 460 | 579 | 415 | 511 | 539 | 770 | 1 | | | | | | |
| 15 | 758 | 556 | 396 | 707 | 335 | 643 | 170 | 598 | 746 | 282 | 763 | 685 | 660 | 713 | 1 | | | | | |
| 16 | 643 | 629 | 545 | 710 | 356 | 496 | 094 | 687 | 731 | 392 | 543 | 604 | 588 | 677 | 788 | 1 | | | | |
| 17 | 054 | 153 | 183 | 154 | 085 | 129 | 059 | 156 | 238 | 120 | -06 | 140 | 215 | 090 | 020 | 250 | 1 | | | |
| 18 | 477 | 586 | 133 | 570 | 220 | 348 | 262 | 494 | 531 | 414 | 433 | 518 | 758 | 790 | 614 | 653 | 210 | 1 | | |
| 19 | 275 | 235 | 262 | 398 | 110 | 108 | 033 | 253 | 313 | 184 | 431 | 362 | 264 | 496 | 375 | 502 | 069 | 464 | 1 | |
| 20 | 336 | 439 | 286 | 405 | 152 | 272 | -16 | 318 | 384 | 240 | 250 | 326 | 410 | 533 | 354 | 525 | 179 | 706 | 538 | 1 |

## Educational contributionand significance of the study.

The scope of the study is addressed in terms of methodological, theoretical, and practical contributions. Methodological contribution can be viewed in terms of evaluation of the effects of item order and context, which have often been assumed to be invariant. Moreover, this study provides an empirical rationale for ordering and grouping items. As for the theoretical contribution, the study demonstrated that item order is a reflection of the cognitive schema and utilization of cues, which are used by respondents. Responses can therefore be modeled by considering the ordering of items and cognitive theories, which account for these responses.

Practical contributions of the study are that variables under consideration should be taken into account when conducting test-equating procedures, and in the construction of Likert-type questionnaires for survey instruments and attitudinal measures. This is because variables in the study are hypothesized to influence item parameter estimates and the accuracy of interpretation of the results.

### Methodological contribution

Methodological contribution of the study is that it provides information and precision on item ordering based on a definite empirical rationale in addition to further evidence of validity to the interpretation of instruments and attitudinal scales. The study provides definite criteria for optimal item ordering of Likert-type questionnaires and survey instruments and demonstrates the effect of item cues among respondents. This necessitates utilization of models that take into account item ordering in order to measure accurately, the targeted attribute. It is from item ordering that hierarchically nested cognitive processes can be inferred. Therefore, findings of this study will have important implications to test interpretations and analysis based on the assumptions of item order and context invariance conditions. In particular, the findings will impact on analysis and valid interpretations of survey instruments.

### Theoretical contribution

The study demonstrated the impact of item order effects, which are a reflection of the schema of the respondents, and in turn provide a link between responses and cognitive processes. Studies on cognition have shown that responses are based on cognitive network or maps from where response cues are drawn. This is evident from cognitive theories such as the information integration theory (Anderson, 1981; Anderson, 1996), and Tatsuoka's rule space model (Tatsuoka & Tatsuoka, 1997).

For sometime test developers have overlooked the fact that responses are based on complex cognitive processes that are dependent on cognitive schema and organization of thought, which are mainly hierarchical. Response to items or task performance dependent on the integration of information gathered from the stimulus. Information Integration theory and other cognitive theories can therefore account for responses and response patterns given the use of cues in cognitive network in the thought schema of the respondents. Mathematical models currently used in test development cannot sufficiently account for response patterns, given the multivariate nature of the response process, and interrelationship of concepts, as well as cues used in responding to items. It is recommended that models that take into account these relationships be incorporated in test designs. The theoretical contribution will be based on the link between the measurement of responses to the cognitive processes that are hierarchicaland sequential.

### Practical contribution

The use of one version of a questionnaire without taking into account the item order effects could have an impact on the test structure, measurement, and interpretation of the targeted construct. Precision with which a construct is measured depends on the optimal ordering of items and the grouping of items. Determination of the optimal order will result in a relatively high information and hence a higher precision in measuring the targeted attribute. Thus, identification of the optimal order will significantly contribute to the test construction in attitudinal scales. Effects of item order and grouping have an impact on test equating procedures, which also assume item order invariance. Test equating procedures will need to be reviewed, taking into account effects of item order and grouping.

This study will have important implications to CAT and MAT where local independence is a basic assumption. The effect of item order on test structure will also have important implications for the gathering of construct related evidence of validity. This will lead to adaptation of appropriate models, which consider item order effects in future test designs and development.

## REFERENCE

Ackerman, T. A. & Spray, J. A. (1986). A general model for item dependency. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA. April 1986.

Ackerman, T.A. (1987). The robustness of the LOGIST and BILOG IRT estimation programs to violation of local independence. Paper presented at the American Educational Research Association. Washington DC. April 1987.

Anderson, N. H. (1981). Foundation of information integration theory. New York: Academic press.

Anderson, N. H. (1996). A functional theory of cognition. New Jersey: Lawrence Erlbaum Associates.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. Journal of Educational Measurement, 22, 3, 231-240.

Chan, J.C. (1991). Response order effects in Likert-type scales. Educational and Psychological measurement, 51, 531-541.

Dorans, N. J., & Lawrence, I.M. (1988). Checking the equivalence of nearly identical test editions. Research report ETS-RR-88-6. College Board Statistical Analysis, ETS. February 1988.

Fishbein, M., & Ajzen, I. (1975). Beliefs, Attitudes, Intention and Behavior: An introduction to the theory and research. Melon Park, CA: Addison-Wesley Publishing Company.

.

Kingston, N. M & Dorans, N. M. (1984). Item location effects and their implication for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.

Krosnick, J. A., & Alwin, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey instrument. Public Opinion Quarterly, 51, 201- 219.

Leary, F. L., & Doran, N.J. (1985). Implications of altering the context in which test items appear: A historical perspective on an immediate concern. Review of Educational Research, 55, 3, 387-413.

Massaro, D. W. & Cowan, N. (1993). Information processing models: Microscopes of the mind. Annual Review of Psychology, 44, 383-425.

Mellenbergh, G. J. (1996). Measurement precision in test score and Item response models. Psychological Methods, 1, 3, 293-299.

Nichols, P. & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. Educational Measurement; Issues and Practice, 2, 18-29.

Reynolds, W. M., Flament, J., Masango, S., & Steele, B. (1999). Reliability and Validity of the Physical Self-Concept Scale. Paper presented at the Annual convention of the American Educational Research Association, Montreal, April, 1999.

Schriesheim, C.A., & Hill, K.D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. Educational and Psychological Measurement, 41, 1101-1114.

Schurr, K.T., & Henriksen, L.W. (1983). Effects of item sequencing and grouping in low- reference type questionnaires. Journal of Educational Measurement, 20, 4, 379-391.

Shavelson, R. J., Hubner, J. J., & Stanton, G.C. (1976). Self-concept: Validation of construct interpretations. Review of Educational Research, 46, 407-441.

Sijtsma, K., & Junker, B.W. (1996). A survey of theory and methods of invariant item ordering. British Journal of Mathematical and Statistical Psychology, 49,79-105.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlets concept in the measurement of Psychopathology. Psychological Methods, 1, 1, 81-97.

Unidimensionality. Psychometrika, 52, 589- 617.

Tatsuoka, M.K., & Tatsuoka, K. K. (1997). Computerized cognitive diagnostic adaptive testing effect on remedial instruction as empirical validation. Journal of Educational Measurement, 34, 1, 3-20.

Yen, W.M. (1984). Effect of local item dependence on the fit and equating performance of the three parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W.M. (1993). Scaling performance assessment: strategies for managing local dependence. Journal of Educational Measurement, 30, 2, 187-213.

Zwick, R. (1991). Effect of item order and context on estimation of NAEP reading proficiency. Educational Measurement: Issues and Practice, 2, 10-15.