European Journal of Psychological Assessment, Vol. 13, Issue 2, pp. 67–74

"The Whole is More than the Sum of its Parts": The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires

Gabriele Helga Franke

University of Essen, Germany

This investigation was concerned with the widely spread practice to extract subscales from extensive psychodiagnostic inventories and to present their items in questionnaires as homogeneous item-blocks. By way of examining the effects of the mode of item-presentation on the reliability and validity, the frequently used and validated SCL-90-R was analyzed as a prototype of multidimensional symptom self-report inventories. Two studies were conducted in different contexts of application (study I: controlled group testing, study II: non-controlled individual testing) and involved different groups of subjects (study I: 130 nursing school students, study II: 134 university students). In both studies the standard item-arrangement was contrasted with the item-block presentation, which groups together items measuring the same dimension. The results revealed significant effects of item-blocking on the mean values, on the reliability, and the validity of the questionnaire. The findings seriously call into question the admissibility of subscale-extraction for self-report inventories. We conclude that the Gestalt paradigm, "The whole is more than the sum of its parts," is valid for multidimensional psychodiagnostic inventories.

Keywords: Item analysis, questionnaires, validity, reliability

Questionnaires are widely used in psychological research and in psychodiagnostic practice. Until now, countless instruments have been developed and employed. The self-report mode of psychological measurement provides safe and exclusive information inaccessible through other channels of assessment. Another advantage of self-report questionnaires is their economical use in professional time.

Today we notice the widely spread practice of extracting subscales from extensive personality inventories or symptom checklists and of presenting their items as homogeneous item-blocks in questionnaires. The main argument advanced in support of this practice is its apparent economic profit. We find this method, for instance, in the application of the Symptom Checklist 90-R (SCL-90-R, Derogatis, 1977, 1992). Several authors select certain dimensions of the SCL-90-R so that they can investigate psychological distress in various populations of patients (Binstock, Foster & Mullane, 1984; Blank & Perry, 1984; Kathol, Noyes, Slymen, Crowe, Clancy & Kerber, 1980; Mellinger, Balter, Manheimer, Cisin & Parry, 1978; Newmann, Engel & Jensen, 1990; Ostrow, Joseph, Kessler, Soucy, Tal, Eller, Chmiel & Phair, 1989; Williams, Rabkin, Remien, Gorman & Ehrhardt, 1991). This method of extracting subscales from extensive psychodiagnostic questionnaires and of presenting the items of these scales in approximately homogeneous item-blocks is almost becoming a standard procedure in all areas of psychological research and practice. Unfortunately, possible methodological implications are not discussed.

For it must be noted that the sequence of the items in a multidimensional questionnaire is the consequence of a decision made during the development of the instrument. This sequence is a characteristic feature of the psychodiagnostic instrument and serves as a basis for the calculation of normative data. Typically, items measuring different constructs are presented to the respondent in a disguised fashion. The underlying assumption is that the purpose of a measurement must be hidden from the respondent, primarily in order to reduce the influence of various response biases such as the desires to look good to others (e. g., evaluation apprehension and social desirability) and to oneself (e. g., cognitive consistency and ego defence mechanisms) (Solomon & Kopelman, 1984). A possible disadvantage of this mode of eliciting responses may lie in the difficulty to give accurate answers to similar questions that are distributed across the whole

questionnaire (Schriesheim, Kopelman & Solomon, 1989).

In the development of an interview it is usually considered a malpractice to neglect effects that may result from the sequence, the context, or possible semantic radiation of questions (Krampen, Hense & Schneider, 1992). It is therefore surprising that the effects of the questionnaire format on its psychometric properties have long been ignored.

Researchers in industrial and organizational psychology studied the possible advantages of arranging items measuring the same constructs into groups and labeling these groups according to their respective scales. Their intention was to counter the monotony of completing long instruments and to encourage the respondents' trust and openness in the research process (Schriesheim & DeNisi, 1980).

Baehr (1953) was the first to test the effects of grouping and randomizing items which measured various job-related attitudes. Obtaining almost identical attitude profiles, she concluded that both modes of measurement were equally accurate. Schriesheim and DeNisi (1980), Schriesheim (1981 a,b), and Solomon and Kopelman (1984) investigated the effects of grouping items, only to conclude later, "These prior studies are not, however, of sufficient quality to allow the drawing of firm conclusions about the effects of item presentation format" (Schriesheim, Kopelman & Solomon, 1989, p. 489). For this reason, Schriesheim, Kopelman and Solomon (1989) explored the effects of grouping and randomizing items on various measures of psychometric quality of the Job Diagnostic Survey (Hackman & Oldham, 1975). They could not find any format to be clearly superior with regard to the examined psychometric properties and concluded that grouping questionnaire items was useless.

In Germany, Krampen, Hense and Schneider (1992) investigated the effects of item-blocking on a German multidimensional personality inventory. Krampen (1993) analyzed the effects of item-blocking on a German inventory that assesses the self-concept, internality, externality to powerful others, and the change of control in locus of control beliefs. In both studies the authors rejected the usefulness of item-blocking because this method was found to influence the mean values, the reliability and the validity of the questionnaires.

The present investigation was aimed at demonstrating that the extraction of subscales from multidimensional psychodiagnostic questionnaires and the presentation of items in homogeneous item-blocks are methodologically unsound. We assume that the Gestalt paradigm "The whole is more than the sum of its parts," also holds for psychodiagnostic inventories. Using the well-known SCL-90-R as a prototype of multidimensional self-report inventories, we conducted two studies in different contexts of application (study I: controlled group testing, study II: non-controlled individual testing) and with different groups of subjects (study I: 130 nursing school students, study II: 134 university students). In both studies the standard item-arrangement (A) was contrasted with the item-blocking presentation, which groups together items belonging to the same original subscale (B).

The analysis of the literature available on this subject suggests the conclusion that item-blocking may lead to unsystematic effects. The semantic homogeneity caused by item-blocking could result in lower mean values because of effects of reactance and defence. On the other hand, it is also possible to homogenize the answers, which would increase internal consistency. For these reasons two-sided statistical tests ought to be done.

Study I

Testing Instruments

We used the SCL-90-R, a multidimensional symptom self-report inventory. It is composed of 90 items, each of which measures distress on a 5-point scale from "not at all" (0) to "extremely" (4). Psychological distress is reflected in nine primary symptom dimensions and three global indices of distress. Historically, the SCL-90-R was developed from the Hopkins Symptom Checklist (HSCL, Derogatis et al., 1974 a,b); today the revised version ("R") is in use. The SCL-90-R quantifies psychological distress in terms of nine primary symptom constructs: Somatization (SOM), Obsessive-Compulsive (O-C), Interpersonal Sensitivity (INT), Depression (DEP), Anxiety (ANX), Hostility (HOS), Phobic Anxiety (PHOB), Paranoid Ideation (PAR) and Psychoticism (PSY). Three additional global indices reflect various aspects of overall psychological distress: GSI (Global Severity Index), PSDI (Positive Symptom Distress Index), and PST (Positive Symptom Total).

We prepared two versions of the SCL-90-R, the standard item-arrangement (A) and the item-blocking arrangement, which groups together the items of each original subscale (B). Both versions were presented in the same professional lay-out to rule out possible side-effects (Champion & Sear, 1968/69).

Method of Analysis

The data were subjected to three analyses. First, two multivariate analyses of variance (MANOVA, Tatsuoka, 1971) were performed. We analyzed the factor "Item Presentation" over the nine dimensions (SOM to PSY). The second multivariate analysis of variance analyzed the factor "Item Presentation" over the three global indices. In preparation for this analysis, the raw scores were transformed to the corresponding T-scores of the German normative sample (Franke, 1995). Secondly, the alpha coefficients of internal consistency were computed for each scale (for the grouped and randomized scores, respectively) and then tested for significant differences. This calculation was done by means of the W-Test (Feldt, 1969; Feldt, Woodruff & Salih, 1987) and took into account the problems of multiple testing (Feild & Armenakis, 1974). Thirdly, the mean intercorrelations were computed for versions A and B as indices of construct validity. Intercorrelations are most economically analyzed by computing the differences of the absolute Fisher-z-transformed (Fisher & Yates, 1963) data.

Subjects and Procedure

The sample of the first study consisted of 134 students (26.1% male, 73.9% female) of the nursing school of the University Hospital Essen, Germany. The sample was subjected to controlled group testing. The standard item-arrangement (A) and the item-blocking arrangements (B) were randomly assigned to the students. The respondents were guaranteed anonymity, and participation was voluntary.

Results

Table 1 shows the results of the two multivariate analyses of variance. Note that the mean T-scores of this sample were typical of a student population (Franke, 1995). Regarding SOM to PSY, there was a significant main effects of the factor "Item Presentation." Subsequent univariate comparisons of the mean values (t-tests) revealed that the factor "Item Presentation" led to significantly lower mean values in the dimensions OB-C, DEP, ANX, PAR, and PSY.

In an analogous MANOVA, we found that the factor "Item Presentation" affected the three global indices, too. Subsequent univariate comparisons again showed that the item-blocking presentation produced significantly lower responses in all three global indices. These global indices are distinct measures of individual distress, reflecting different aspects of psychopathology (Derogatis, Yvzeroff & Wittelsberger, 1975). Item-blocking reduced the number of symptoms reported by the participant (PST), decreased the parameter indicating intensity of response (PSDI, "corrected" for number of symptoms), and led to a lower global measure of overall distress (GSI).

Table 1. Study I: Basic data of the SCL-90-R dimensions and global indices in a group of 134 nursing school students. Condition A (standard item arrangement, N = 67), Group B (items measuring the same dimension grouped together, N = 67).

SCL-90-R subscales	Group A x	s	Group B x	s	*t*-test t	*p* <
SOM	53.3	12.6	54.3	7.4	0.6	.55
OB-C	57.5	10.4	51.0	9.0	3.9	.00
INT	58.1	10.7	57.0	8.0	0.7	.47
DEP	56.6	11.0	49.8	7.6	4.1	.00
ANX	55.0	10.0	49.8	9.1	3.2	.01
HOS	56.3	10.8	53.2	8.2	1.9	.06
PHOB	52.5	9.4	51.8	10.6	0.4	.67
PAR	52.3	10.9	52.2	10.5	3.3	.00
PSY	53.8	10.3	47.8	8.1	3.7	.00
Analysis of variance SOM to PSY (MANOVA)						
F(9/124)	Significance					
5.86	.000					
SCL-90-R Global indices	Group A x	s	Group B x	s	t-test t	p <
GSI	57.3	11.6	52.4	7.4	2.9	.01
PSDI	56.9	11.7	53.2	8.0	2.1	.04
PST	56.7	11.8	52.1	7.8	2.7	.01
Analysis of variance GSI, PSDI, PST (MANOVA)						
F(3/130)	Significance					
2.85	.040					

Table 2. Study I: Reliability of the SCL-90-R dimensions in a group of 134 nursing school students. Condition A (standard item arrangement, N = 67), Group B (items measuring the same dimension grouped together, N = 67).

SCL-90-R subscales and GSI	Group A alpha	Group B alpha	W-Test F(66/66)
SOM	.79	.46	2.57*
O-C	.86	.65	2.51*
INT	.83	.78	1.32
DEP	.90	.67	3.13*
ANX	.88	.78	1.79+
HOS	.80	.65	1.80+
PHOB	.66	.87	2.51*
PAR	.72	.76	1.14
PSY	.78	.59	1.89*
GSI	.97	.92	2.75*

+ = p < .05, * = p < .01

Table 3. Validity measures in Study I: Intercorrelations of the nine subscales.

Scale		Standard Item Arrangement (Group A)						
	SOM	O-C	INT	DEP	ANX	HOS	PHOB	PAR
O-C	.45							
INT	.33	.78						
DEP	.51	.80	.67					
ANX	.49	.74	.56	.73				
HOS	.26	.69	.65	.57	.57			
PHOB	.46	.65	.51	.64	.75	.58		
PAR	.32	.63	.76	.57	.56	.57	.56	
PSY	.50	.65	.67	.56	.68	.45	.51	.69
Scale		Items Measuring the Same Dimension Grouped Together (Group B)						
	SOM	O-C	INT	DEP	ANX	HOS	PHOB	PAR
O-C	.28							
INT	.35	.62						
DEP	.26	.65	.59					
ANX	.56	.49	.46	.44				
HOS	.24	.51	.43	.43	.45			
PHOB	-.01	-.28	-.11	.06	.25	-.10		
PAR	.30	.53	.69	.53	.59	.30	-.04	
PSY	.45	.60	.42	.47	.42	.24	-.08	.45

	Condition		
	Group A	Group B	
			d(z)A,B
x	.59	.38	
z	.69	.42	.27+

+ = $p < .05$

For each of the grouped and randomized measurements, the internal-consistency reliabilities (Cronbach's α) were separately computed for each scale and for the global index GSI. The obtained values were then tested for significant differences (Table 2).

The standard item arrangement led to parameters of reliability similar to those of the normative sample. The relatively low value of the coefficient alpha in the subscale PHOB is only found in healthy but not in clinical populations (Franke, 1995). The W-Test (Feldt, 1969; Feldt, Woodruff & Salih, 1987) detected significant differences between the two conditions of the factor "Item Presentation" in eight comparisons. Seen against a total of ten comparisons, this cannot be considered a chance result ($p < .0001$, Feild & Armenakis, 1974). Item-blocking led to lower reliability – except for the scale PHOB, which had a higher coefficient alpha. Thus, the extraction of items of subscales from the extensive psychodiagnostic inventory and their presentation in a blocked way caused unsystematic but significant changes in the parameters measuring the reliability of the scales.

The matrix of the intercorrelations of the scales was analyzed as an aspect of construct validity (Table 3). Table 3 also reports the differences of the absolute Fisher z-transformed (Fisher & Yates, 1963) coefficients of intercorrelation. The standard item arrangement led to parameters of intercorrelation similar to a student population (Franke, 1995), whereas the item-blocking arrangement resulted in an unusual intercorrelation matrix with several unexpected intercorrelations and in a significantly lower absolute mean intercorrelation.

Item-blocking caused significantly lower mean values as well as significant changes in the parameters of reliability. It also produced an unusual intercorrelation matrix as an indicator of construct validity. In other words item-blocking impaired the differentiation of the constructs underlying the test.

Study II

Measure and Method of Analysis

In order to replicate the findings of study I, study II used the same two forms of the SCL-90-R and the same methods of analysis.

Table 4. Study II: Basic data of the SCL-90-R dimensions and global indices in a group of 130 university students. Condition A (standard item arrangement, N = 65), Group B (items measuring the same dimension grouped together, N = 65).

SCL-90-R subscales	Group A x	Group A s	Group B x	Group B s	t-test t	t-test p <
SOM	54.1	10.4	52.7	11.3	0.7	.46
O-C	57.7	9.9	50.8	9.3	4.1	.00
INT	56.1	13.4	51.8	7.9	2.2	.03
DEP	56.9	12.1	50.5	9.4	3.4	.00
ANX	60.1	9.5	53.2	10.1	4.1	.00
HOS	55.9	10.2	49.9	11.7	3.2	.00
PHOB	54.5	11.6	60.1	9.7	3.0	.00
PAR	53.8	10.9	51.6	10.8	1.2	.26
PSY	56.3	13.6	57.7	9.5	0.7	.49

Analysis of variance SOM to PSY (MANOVA)

F(9/120)	Significance
8.41	.000

SCL-90-R Global indices	Group A x	Group A s	Group B x	Group B s	t-test t	t-test p <
GSI	58.3	11.6	54.1	9.6	2.3	.03
PSDI	58.4	8.6	56.1	8.6	1.5	.13
PST	57.6	12.0	52.6	10.0	2.6	.01

Analysis of variance GSI, PSDI, PST (MANOVA)

F(3/126)	Significance
3.30	.022

Subjects and Procedure

The sample for the second study consisted of 130 undergraduate students (48% male, 52% female) of the University Hospital Essen, Germany. The sample was subjected to non-controlled individual testing. As in study I, the standard item-arrangement (A) and the item-blocking arrangement (B) were randomly assigned to the students, who participated voluntarily and were guaranteed anonymity.

Table 5. Study II: Reliability of the SCL-90-R dimensions in a group of 130 university students. Condition A (standard item arrangement, N = 65), Group B (items measuring the same dimension grouped together, N = 65).

SCL-90-R subscales and GSI	Group A alpha	Group B alpha	W-Test F(64/64)
SOM	.61	.69	1.24
O-C	.68	.70	1.06
INT	.87	.54	3.66*
DEP	.88	.82	1.56+
ANX	.71	.61	1.34
HOS	.58	.55	1.06
PHOB	.51	.46	1.10
PAR	.75	.67	1.30
PSY	.80	.53	2.39*
GSI	.96	.93	1.78+

+ = p < .05, * = p < .01

Table 6. Validity measures in Study II: Intercorrelations of the nine subscales.

Scale	Standard Item Arrangement (Group A)						
	SOM	O-C	INT	DEP	ANX	HOS	PHOB	PAR
O-C	.41							
INT	.55	.67						
DEP	.58	.65	.85					
ANX	.57	.65	.67	.78				
HOS	.44	.39	.52	.55	.28			
PHOB	.52	.52	.74	.66	.56	.39		
PAR	.53	.60	.85	.79	.59	.58	.77	
PSY	.56	.55	.84	.78	.74	.46	.67	.81

Scale	Items Measuring the Same Dimension Grouped Together (Group B)						
	SOM	O-C	INT	DEP	ANX	HOS	PHOB	PAR
O-C	.38							
INT	.47	.69						
DEP	.46	.62	.68					
ANX	.38	.29	.40	.45				
HOS	.39	.54	.60	.60	.44			
PHOB	.26	.19	.22	.18	.40	-.01		
PAR	.14	.45	.47	.42	.40	.47	.04	
PSY	.41	.39	.42	.46	.56	.36	.41	.32

	Condition Group A	Group B	d(z)A,B
x	.61	.40	
z	.75	.44	.31 +

+ = p < .05

Results

For brevity's sake, the results will be presented in the above manner.

Table 4 shows the results of the multivariate analysis of variance. The mean T-scores of those subjects who had been assigned the standard item-arrangement were typical of a student population (Franke, 1995). There was a significant main effect of the factor "Item Presentation." Subsequent univariate comparisons (t-tests) showed that the factor "Item Presentation" led to significantly lower mean values in the dimensions OB-C, INT, DEP, ANX, HOS and to a significantly higher mean value in the dimension PHOB. In an analogous MANOVA of the three global indices, the factor "Item Presentation" was again found to have a main effect. The univariate comparisons of the mean values disclosed that the item-blocking resulted in significantly lower responses in GSI and PST. This means that item-blocking brought about different results that corresponded to those of study I presented above.

Table 5 presents the results of the analysis of the reliability parameters. Item-blocking produced significantly lower values for the reliability parameters in three scales, and in GSI. These results again differed from those of study I.

For the standard item arrangement we found usual parameters of intercorrelation. The item-blocking arrangement again led to an unusual intercorrelation matrix and to a significantly lower mean intercorrelation (Table 6).

Discussion and Conclusion

In order to evaluate the effects of item-blocking – that is, of extracting subscales from a more extensive psychodiagnostic inventory and presenting the items in homogeneous groups – on the mean values and the reliability and validity parameters of a questionnaire, the widely used SCL-90-R was analyzed as a prototype of multidimensional self-report inventories. Two studies were conducted in different contexts of application (study I: controlled group testing, study II: non-controlled individual testing) and involved different groups of subjects (study I: 130 nursing school students, study II: 134 university students). In both studies the standard item-arrangement (A) was contrasted with the item-block presentation, in which items measuring the same dimension were grouped together (B). The results showed that item-blocking significantly affects the mean values, the reliability, and the validity of the questionnaire.

The high quality of both studies is evidenced by the fact that the two groups completing the original version of the SCL-90-R did not differ in their mean values and reliability and validity parameters from comparable normative data (Franke, 1995). The significant factor "Item Presentation" produced lower mean values.

We also detected significant effects of item-blocking on the internal-consistency reliabilities as well as on the intercorrelation of the subscales. Therefore, the original version of the SCL-90-R and the item-blocking version are not empirically equivalent.

Apart from the essential correspondence of the main results of the two studies, there were several differences between individual results. Posterior t-tests of mean differences showed few correspondences between the two studies. Study I and II also differed concerning those scales in which differences of internal consistency were detected after item-blocking. Thus neither the general hypothesis that item-blocking completely homogenizes the response-behavior nor the hypothesis that reactance is increased by the repetition of semantically similar questions could be confirmed. The significant effects were unsystematic. Since the intercorrelation of the scales as an indicator of construct validity differs from normative data when items are blocked, this mode of presentation evidently makes more difficult the differentiation of constructs.

The final score does depend on item order. In other words: the SCL-90-R items show item dependency to their original sequence. This sequence is the characteristic feature of the psychodiagnostic instrument and serves as the basis of the calculation of normative data. It seems impossible to change this characteristic item sequence without measurable effects on the means, the reliability and validity parameters. In the case of the SCL-90-R, the purpose of the measurement is hidden from the respondent by the original item sequence. The item-blocking mode, on the other hand, allows the subject to infer it. Inferring the purpose of the measurement, however, leads to unusual response behavior, which cannot be explained by usual response bias theories. In light of our results, it may be interesting to study the effects of other semantically meaningful and senseless (random) item sequences. Perhaps this investigation may help to answer the question about item dependency.

The findings of this investigation corroborate our hypothesis that the extraction of SCL-90-subscales

is a malpractice. The results of our study suggest that the altered scale for computing T-scores should not be applied in any case (individual diagnosis, group evaluation, validation of another scale, etc.). In view of the low internal consistencies of some altered versions their use is highly questionable anyway.

Both investigated samples differ in some important respects: in the level of education (study I lower level, study II higher level), in the distribution of gender (study I 73.9% women, study II 52% women), and in the contexts of application (study I: group testing, study II: individual testing). The T-transformation should minimize the well-known effects of the level of education and of gender on the scores of the SCL-90-R. Nonetheless, it cannot be ruled out that gender or the educational level or the context of application contributes to the differences between our studies. Regarding the results of McKelvie (1986), who found gender differences in an investigation of the Vividness of Visual Imagery Questionnaire, it should be an aim of future investigations to look for similar effects in comparably large groups.

Finally, after rejecting item-extraction and item-blocking as methodological malpractices on the evidence presented here, we are still confronted with the question of what to do when a multidimensional personality inventory or a symptom checklist is found to be too long. In the case of the SCL-90-R this question is easily answered because a short version, the Brief Symptom Inventory (BSI, Derogatis & Melisaratos, 1983; Derogatis, 1993) of only 53 items, is available. Generally speaking, however, an exact consideration of the methodological problems involved is a prerequisite to a purposeful choice from the multitude of available psychodiagnostic inventories.

Our investigation has demonstrated that the extraction of subscales from self-report inventories is a highly doubtful method because the reliability and validity parameters of the subscales change markedly. Therefore we conclude that the Gestalt paradigm "The whole is more than the sum of its parts," finds a modern application in psychodiagnostic inventories. Therefore, if a multidimensional psychodiagnostic inventory is divided into its subdimensional parts, and if these parts are then presented in isolation and grouped into blocks, one will get an entirely new inventory rather than a part of the original one. This is because the whole is indeed more than the sum of its parts.

Acknowledgments

We would like to thank Günther Krampen for his insightful comments on earlier studies, Gisla Gniech for her valuable reference to the Gestalt paradigm, Ralf Baethke-Franke for his help in preparing the two versions of the questionnaire, Uta Schneider and Hellmut Hasenburg for their assistance. Finally, we are grateful to all students who participated in these studies.

Author's address

Gabriele Helga Franke
Institute of Medical Psychology
University Hospital
Hufelandstrasse 55
45122 Essen
Germany

References

Baehr, M.E. (1953). A simplified procedure for the measurement of employee attitudes. *Journal of Applied Psychology, 37,* 163–167.

Binstock, G., Foster, L.G., & Mullane, J.F. (1984). Propanolol and the depression component of anxiety neurosis. *Current Therapeutic Research, 35,* 423–432.

Blank, K., & Perry, S. (1984). Relationship of psychological processes during delirium to outcome. *American Journal of Psychiatry, 141,* 843–847.

Champion, D.J., & Sear, A.M. (1968/1969). Questionnaire response rate: A methodological analysis. *Social Forces,* 335–339.

Derogatis, L.R. (1977). *SCL-90-R, administration, scoring, & procedures manual-I for the R(evised) version.* Johns Hopkins University School of Medicine.

Derogatis, L.R., Lipman, R.S., Rickels, K., Uhlenhuth, E.H., & Covi, L. (1974a). The Hopkins Symptom Checklist (HSCL). A measure of primary symptom dimensions. In Pichot, P. (Ed.), *Psychological measurements in psychopharmacology* (pp. 79–110). Basel: Karger.

Derogatis, L.R., Lipman, R.S., Rickels, K., Uhlenhuth, E.H., & Covi, L. (1974b). The Hopkins Symptom Checklist (HSCL): a self-report symptom inventory. *Behavioral Science, 19,* 1–13.

Derogatis, L.R., Yvzeroff, H., & Wittelsberger, B. (1975). Social class, psychological disorder, and the nature of the psychopathologic indicator. *Journal of Consulting and Clinical Psychology, 43,* 183–191.

Derogatis, L.R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine, 13,* 595–605.

Derogatis, L.R. (1992). *SCL-90-R, administration, scoring, & procedures manual-II for the R(evised) version and other instruments of the Psychopathology Rating Scale Series.* Townson: Clinical Psychometric Research, Inc.

Derogatis, L.R. (1993). *BSI, administration, scoring, and*

procedures manual, Third edition. Minneapolis: National Computer Systems, Inc.

Feild, H.S., & Armenakis, A.A. (1974). On use of multiple tests of significance in psychological research. *Psychological Reports, 35,* 427–431.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's α or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34,* 363–373.

Feldt, L.S., Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient α. *Applied Psychological Measurement, 11,* 93–103.

Fisher, R.A., & Yates, F. (1963). *Statistical tables for biological, agricultural and medical research*. London: Oliver and Boyd.

Franke, G.H. (1995). *SCL-90-R. Die Symptom-Checkliste von Derogatis – Deutsche Version*. Göttingen: Beltz Test.

Hackman, J.R., & Oldham, G.R. (1975). Development of the job diagnostic survey. *Journal of Applied Psychology, 60,* 159–179.

Kathol, R.G., Noyes, R., Slymen, D.J., Crowe, R.R., Clancy, J., & Kerber, R.E. (1980). Propanolol in chronic anxiety disorders. *Archives of General Psychiatry, 37,* 1361–1365.

Krampen, G. (1993). Effekte von Bewerbungsinstruktionen und Subskalenextraktion in der Fragebogendiagnostik. *Diagnostica, 39,* 97–108.

Krampen, G., Hense, H., & Schneider, J.F. (1992). Reliabilität und Validität von Fragebogenskalen bei Standardreihenfolgen versus inhaltshomogener Blockbildung ihrer Items. *Zeitschrift für experimentelle und angewandte Psychologie, 34,* 229–248.

McKelvie, S.J. (1986). Investigation of the Vividness of Visual Imagery Questionnaire. *British Journal of Psychology, 77,* 229–36.

Mellinger, G.D., Balter, M.B., Manheimer, D.I., Cisin, I.H., & Parry, H.J. (1978). Psychic distress, life crisis, and use of psychotherapeutic medications. *Archives of the General Psychiatry, 35,* 1045–1052.

Newmann, J.P., Engel, R.J., & Jensen, J. (1990). Depressive symptom patterns among older women. *Psychology and Aging, 5,* 101–118.

Ostrow, D.G., Joseph, J.G., Kessler, R., Soucy, J., Tal, M., Eller, M., Chmiel, J., & Phair, J.P. (1989). Disclosure of HIV antibody status: Behavioral and mental health correlations. *AIDS Education and Prevention, 1,* 1–11.

Schriesheim, C.A. (1981 a). The effect of grouping or randomizing items on leniency response bias. *Educational and Psychological Measurement, 41,* 401–411.

Schriesheim, C.A. (1981 b). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. *Educational and Psychological Measurement, 41,* 1093–1099.

Schriesheim, C.A., & DeNisi, A.S. (1980). Item presentation as an influence on questionnaire validity: A field experiment. *Educational and Psychological Measurement, 40,* 175–182.

Schriesheim, C.A., Kopelman, R.E., & Solomon, E. (1989). The effect of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. *Educational and Psychological Measurement, 49,* 487–508.

Solomon, E., & Kopelman, R.E. (1984). Questionnaire format and scale reliability: An examination of three modes of item presentation. *Psychological Reports, 54,* 447–452.

Tatsuoka, M.M. (1971). *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley.

Williams, J.B.W., Rabkin, J.G., Remien, R.H., Gorman, J.M., & Ehrhardt, A.A. (1991). Multidisciplinary baseline assessment of homosexual men with and without Human Immunodeficiency Virus Infection. *Archives of the General Psychiatry, 48,* 124–130.