Review of Psychology
1999, Vol. 6, No. 1-2, 25-31 UDC 159.9

Multiple moderator effects on a testee's answer to personality questionnaire items

KLAUS D. KUBINGER, THOMAS KARNER and STEFAN MENGHIN

This paper gives the results of a pilot study and a hypotheses-testing experiment concerning several moderator effects on answering a personality questionnaire. All above, the effect of individual vs. group administration was of interest, the effect of item response format, and the effect of computer vs. paper-pencil administration. All these effects in interaction and in dependence of sex, respectively, gained plenty of interest as well. The experiment tried particularly to test whether or not testees give to a computer rather than to a physical investigator valid answers concerning private and intimate items.

Psychometric analysis, that is, the application of the *Rasch* model, based on 338 and 836 testees of the pilot study demonstrates that the MBTI scales (*Myers-Briggs Type Indicator*) largely fulfil the standards if the testees were tested by computer and a continuous item response format was used; but they do not at all if the testees were tested by paper and pencil and a dichotomous item response format was used. This fact might happen according to the well-known reactancy phenomenon, that is, because almost no latitude testees behave arbitrarily and erratically.

When the MBTI was enriched by a 12 item scale of „taboo" themes, four-way layout multivariate analysis based on 390 subjects resulted as follows: all but one main effect, i.e., computer administration, turned out to be significant, i.e., sex, item response format, and individual vs. group administration. Only in interaction as with sex so with dichotomous vs. six-categorical item response format computer administration established significant effects - however, this is true just for the taboo scale but not for any MBTI scale! It looks like female testees answer taboo items to a computer more male-like than they do by paper-pencil. And the cited reactancy phenomenon seems to be avoidable only by computer administration. Finally, as regards individual vs. group administration, testees surrounded with others claim not as much tabooed behavior as testees do at individual administration, and achieve higher extraversion scores.

All these discribed moderator effects give rise to abandon the idea that personality questionnaires pay. However, for testing a testee's personality an alternative concept is indicated.

This paper does not pursue the well-known skepticism concerning a testee's profit-aimed manipulation of answers given to a personality questionnaire; for this, another paper will be given (but see for the while Kubinger, 1996). This paper deals with the almost unconscious effects that occur while a questionnaire is being answered, and with some psychological or ecological moderator effects.

Mainly the following features are of interest: Computer administration, group administration, item response format, and sex. And the following questions arise:

1) Does individual administration of a personality questionnaire produce other scores than group administration?

2) Do dichotomous item response format and some multi-categorical item response formats result in corresponding interpretations of a testee's traits ?

3) Does it make any difference for the scores whether a testee is tested with or without an investigator present?

4) Has administration by computer any influence on the score, and, in particular, does the computer encourage a testee to give more valid answers to private and intimate items ?

5) Do any of these effects depend on the testee's sex ?

As there are no or just a few ambiguous empirical results that answer these questions this paper offers results and some general conclusions based on two studies: a pilot study and a hypotheses-proving experiment.

Nevertheless, the few results concerning a computer's effect on personality questionnaire scores must be mentioned.

Klaus D. Kubinger, Department of Psychology, Division of Psychological Diagnostics, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria. (Correspondence concernig this article should be sent to this address). Thomas Karner & Stefan Menghin, Department of Psychology, Division of Psychological Diagnostics, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria.

First of all, the results of two recently published studies on psychometric equivalence of a test up-date the findings of former studies: While Bader, Hofmann and Kubinger (1993) proved computer and paper-pencil issue of the German-made questionnaire *Gießen-Test* to be equivalent - i.e., there were neither differences in means and standard deviations nor was there any item bias - Schwenkmezger and Hank (1993) discovered for both state scales of *Spielberger's* STAI and STAXI significant differences, i.e., the computer enhances both actualized anxiety and actualized anger. The consequence of these contradicting empirical results is that every computerization of a questionnaire must be checked for equivalence before it is used in practice. This consequence is however not of much interest in this paper. On the other hand, some non-significant differences do hardly explain whether or not the computer encourages a testee to give more valid answers, particularly to private and intimate items: perhaps the *Gießen-Test's* items do not invade intimate sphere.

Secondly, some papers indicate in fact a higher frankness of testees in front of a computer than in front of a physical investigator: Evan and Miller (1969); Koson, Kittchen, Kochen and Stodolsky (1970); O'Brian and Dudgale (1978) - however, this papers were done obviously in a „prehistoric" PC-age.

METHODS

For a typical personality questionnaire we decided on MBTI (*Myers-Briggs Type Indicator*), German version, because it is a test used world-wide for personality assessment and, nevertheless, does not run the risk that the subjects are acquainted with it. The MBTI is based on *C.G. Jung's* typology and has been modified by *Katherine Briggs* and *Isabel Myers*. There are four bipolar scales - with a total of 90 items in the German version: *extraversion vs. introversion, sensing vs. intuition,* that is, the way of a person's perceptions, *thinking vs. feeling,* that is, the way of a person's evaluations, and *judging vs. perceiving,* that is the way of a person's habits. The four scales with two poles each result in a combination of 16 types, that is, *extraversion-sensing-thinking-judging* or *introversion-sensing-thinking-judging or extraversion-intuition-thinking-judging* and so on up to *introversion-intution-feeling-perceiving.* Though, every scale also offers a simple score. One item of *thinking vs. feeling* is for example:

„Are you more careful about

A people's feelings, or

B their rights?"

- that is, the item response format is originally dichotomous.

The pilot study

The following study was done to test MBTI's psychometric qualification in general and within computer administration in particular, and to answer questions 2) and 3).

338 subjects were tested by the computerized MBTI, the item response format of which was changed into a continuous rating scale. And 836 subjects were tested by the original MBTI paper-pencil edition. The first group primarily consisted of students of psychology, the second of middle- and lower-echelon managers. While to the first group the questionnaire was administered in the standard way (and individually), it was sent by mail to the managers. Of course, the postal „administration" resulted in (anonymous and) voluntarily given answers[1]; in contrast, most of the students had to go through the test - no matter by which score - because of the university curriculum.

Psychometric analysis according to item response theory demonstrates the following.

If *Müller's model for continuous rating scales* (cf. Kubinger, 1989) is applied to the computerized MBTI it proves valid after few items are deleted: i.e., every scale measures unidimensionally and the respective scores are sufficient statistics of subjects' trait parameters. As for this model no (conditional) likelihood-ratio test exists, see as an example Figure 1 for the graphic model check of the scale *extraversion vs. introversion.* While in Figure 1 item parameters estimated for the high scorers are opposed to those estimated for the low scorers two other graphic model checks were done with these opposing estimates based on female vs. male and older vs. younger testees. As far as it concerns the other scales the graphics look alike.

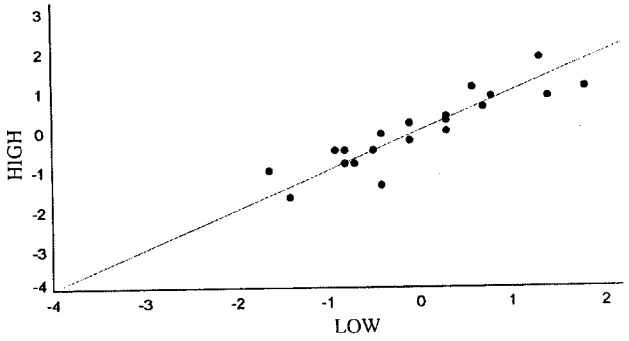

Figure 1. Graphic model check of *Müller's model for continuous rating scales*: MBTI scale *extraversion vs. introversion* after deletion of 7 items.

[1] Thanks to Monika Fabian for organizing this sampling of data

As a matter of fact, if the well-known *Rasch model* is applied after dichotomizing the continuous ratings the results are almost the same: the model is valid for every MBTI scale if few items are deleted. See the pertinent model checks and their statistics in Table 1, and see again as an example the graphic model check of the scale *extraversion vs. introversion*, in Figure 2. Although the items to be deleted are not exactly the same for the two models we may conclude that these analyses show on the whole MBTI's psychometric qualification. Of course, *Müller's model* deliberately takes more information into account, which may mean more reliable scores, but dichotomous as well as continuous scoring leads to psychometrically founded estimates of the trait parameter in question. Subsequently, the model-check-based shortened version (according to the *Rasch model*) will be referred to as „MBTI-S".

Keep in mind that these are the results concerning (a) the computerized, (b) continuous rating scaled MBTI which (c) was administered to testees who were primarily students, who (d) were not really volunteers, but (e) were frequently surrounded by other testees, and where (f) an investigator was present.

Analyzing the data of the original MBTI leads, however, to absolutely contradictory results: the *Rasch model* failed every time, that is, no deletion of items succeeds; at the end only a few items seem to fit the model. Table 2 shows the pertinent model checks and their statistics, and Figure 3 shows the best of two graphic model checks of any scale - if all items were taken into account.

Table 1

Anderson's likelihood-ratio test applied on *Rasch*-analyzed MBTI scales (Chi-square distributed statistics; level of significance: (α=.05) - dichotomized continuous rating data.

	high vs. low	female vs. male	older vs. younger
extraversion vs. introversion	25.89 (df=19) p<0.05	29.91 (df=19) p<0.05	20.24 (df=19) p<0.05
sensing vs. intuition	21.55 (df=14) p<0.05	20.68 (df=14) p<0.05	21.12 (df=14) p<0.05
thinking vs. feeling	14.46 (df=11) p<0.05	17.32 (df=11) p<0.05	11.45 (df=11) p<0.05
judging vs. perceiving	10.89 (df=13) p<0.05	20.70 (df=13) p<0.05	16.58 (df=13) p<0.05

Table 2

Anderson's likelihood-ratio test applied on *Rasch*-analyzed MBTI scales (Chi-square distributed statistics; level of significance: (α =.05) - dichotomous item response format data.

	high vs. low	older vs. younger
extraversion vs. introversion	191.76 (df=26) p<.05	103.39 (df=26) p<.05
sensing vs. intuition	92.49 (df=20) p<.05	52.18 (df=20) p<.05
thinking vs. feeling	533.2 (df=20) p<.05	83.12 (df=20) p<.05
judging vs. perceiving	118.81 (df=19) p<.05	77.57 (df=19) p<.05

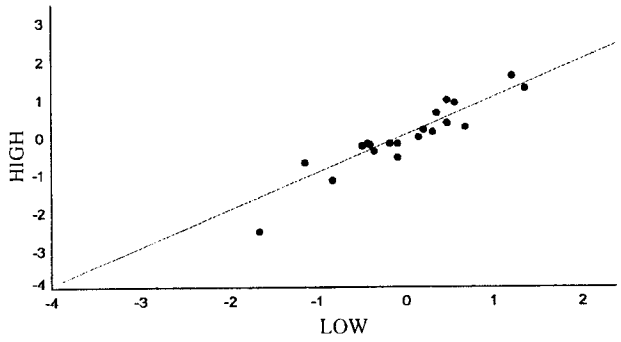

Figure 2. Graphic model check of the *Rasch model:* MBTI scale *extraversion vs. introversion* after deletion of 7 items - dichotomized continuous rating data.

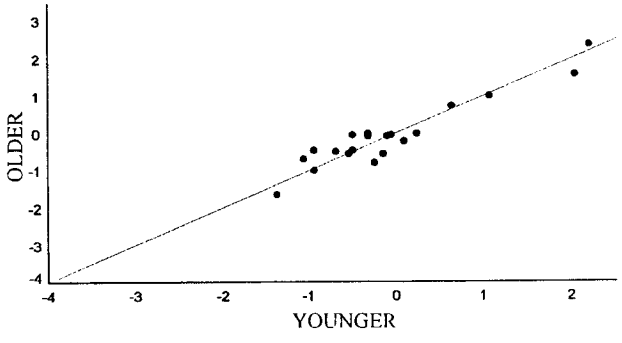

Figure 3. Graphic model check of the *Rasch model:* MBTI scale *extraversion vs. introversion* - dichotomous item response format data.

As these results concern (a) paper-pencil administration using (b) the original dichotomous item response format, and concern (c) testees being subjects of the target population who (d) were not only volunteers but engaged in making public their „performances", who (e) were not surrounded by other testees, and (f) not even faced with the presence of an investigator, the interpretation of contradiction is unequivocal. There is no evidence which of these variations are responsible for.

However, we may for a while conclude that circumstance (b) is the trouble-maker: Indeed, we do not reject the hypothesis that any or even all other circumstances are responsible for the model failing to fit the data, but we got the impression that the dichotomous item response format is to blame, at any rate. According to the theory of psychological reactance (cf. Brehm, 1966) the dichotomous item response format results in a situation that does not allow testees' behavior a great latitude, and as a consequence a testee may behave arbitrarily, erratically, and not at all in conformity with its traits.

Anyhow, we must take into account that perhaps paper-pencil administration causes the model rejection, either due to the phenomenon of social desirability or due to testees' shyness to give valid answers to private and intimate items.

A hypotheses testing experiment

In order to separate all yet indicated moderator effects on the score of a personality questionnaire an experimental approach was applied. The design was based on the following:

- MBTI and MBTI-S[2], respectively, should be supplemented by a scale of „taboo" themes, because the hypothesis that the computer enhances the frankness of testees may be true only if the items get to the bottom of the intimate sphere.

- Individual and group administration as well as paper-pencil and computer administration should be balanced; in particular, group administration should take place at the computer, too.

- Dichotomous and multi-categorical item response format should be opposed; for reasons of dichotomization an even number of (multi-) categories is to be preferred.

- The testees' sex should be uniformly distributed with respect to every combination of experimental factors.

[2] In order to use an almost equal number of items for each scale, 5 items more than necessary due to the *Rasch model* were deleted in the scale *extraversion vs. introversion*; hence, the MBTI-S consists of 15 + 14 + 12 + 14 = 55 items.

	C		P	
	2 cat	6 cat	2 cat	6 cat
	m f	m f	m f	m f
Group		1 . . . 73		74 . . . 136
Individual	137 . . . 207	208 . . . 269	270 . . . 330	331 . . . 390

Figure 4. The four-way layout used.

Thus a four-way layout applies: (individual vs. group administration; factor IG) x (computer vs. paper-pencil administration; factor CP) x (dichotomous vs. multi-categorical item response format; factor DM) x (sex; factor SEX). For organizational reasons a total cross-classification did not occur, that is the combination dichotomous item response format at group administration is missing as at paper-pencil so at computer administration. Figure 4 illustrates the design.

Pre-tests lead to the following items on the intended *taboo scale* (if to be answered dichotomously then either by „rather right" or „rather wrong"):

1 „If some relative dies, I think about inheriting some parts of his/her property"

2 „I have betrayed somebody because of a trifle so that my own mistakes remain undetected"

3 „When my eyes fall on a handicapped person I look away with shame"

4 „Should I ever have problems with money, I would tell my co-workers"

5 „I sometimes dream of an erotic night with my favorite movie star"

6 „I love sucking parts of the body"

7 „If a (married) couple I know break up, I avoid seeing at least one of them"

8 „I have never felt the urge to enter a sex shop"

9 „I spend more time than necessary in the toilet"

10 „I have practised masturbation"

11 „Pictures of well-built men and/or women trigger erotic feelings in me"

12 „I regularly wash my genitals"[3]

[3] For scoring, the last item and item 8 must be reversed.

The intention of these taboo items was not to ask for faults almost everybody has. But it was intended to use taboos that make a testee rather deny than admit them, if the behavior in question applies.

Keep in mind that the scoring rules are not yet based on psychometric analyses, and that therefore item-wise analyses must be added to the analysis of scale scores.

With the multi-categorical item response format, the testees had to differentiate between two extreme answers according to „<<<„, „<<„, „<„, „>„, „>>„ and „>>>„, that is, six-categorically - this is true for MBTI-S and the *taboo scale* as well. Again for scoring, the answers were dichotomized at the end.

Within the given design a total of 390 subjects were randomized. They were students, primarily of psychology, and university graduates. There were 185 male and 205 female subjects, aged between 18 and 62 (M=24.4, SD=5.1, Mdn=23.0).

Multivariate analysis of variance was applied. All four factors were fixed. As these factors were not completely combined (see Fig. 4) the four-way interaction as well as one of the two-way interactions (DMxIG) and two three-way interactions (DMxIGxCP, DMxIGxSEX) cannot be estimated.

RESULTS

The MBTI-scale *sensing vs. intuition* proved to cause heterogeneity of the variance-covariance matrix (*Box's M-test:* $F(165/\infty)=1.41$, $p<0001$); even univariate analysis failed for this scale because of heterogeneity of variance (*Bartlett*-test: $F(11/\infty)=2.35$, $p=.007$). This scale was therefore excluded from further analyses.

Table 3

Main and interaction effects according to multivariate analysis for a four-way layout (*taboo scale* and three MBTI scales).

effect	*Wilk*'s lambda	significance
CP x IG x SEX	.98784	p =.331
CP x DM x SEX	.97986	p =.105
IG x SEX	.99898	p =.984
DM x SEX	.99760	p =.924
CP x SEX	.96242	p =.006
CP x IG	.98656	p =.279
CP x DM	.97454	p =.046
SEX	.89271	p =.000
IG	.95811	p =.003
DM	.97204	p =.031
CP	.99794	p =.942

Table 4

F-statistics of univariate four-way layout analysis of variance (df=1/378) given the overall effect is significant.

	taboo	*extraversion*	*thinking*	*judging*
CP x SEX	9.67 (p=**.002**)	.21 (p=.651)	3.65 (p=.057)	1.36 (p=.244)
CP x DM	5.67 (p=**.018**)	.00 (p=.982)	1.36 (p=.243)	2.06 (p=.152)
SEX	22.37 (p=**.000**)	2.53 (p=.113)	17.42 (p=**.000**)	.00 (p=.976)
IG	5.09 (p=**.025**)	5.65 (p=**.018**)	3.68 (p=.056)	.15 (p=.697)
DM	7.85 (p=**.005**)	3.74 (p=.054)	.06 (p=.801)	.58 (p=.446)

Table 5

Considerable differences of cell-means. High scores in the *taboo scale* indicate a testee's disposition to claim some tabooed behavior. High scores in any MBTI scale indicates a well-characterized testee with respect to the measured trait.

	taboo scale			
	SEX		DM	
	f	m	dichotom.	six-categ.
Computer	4.07	4.67	3.9	4.59
Pencil	3.19	4.7	3.93	3.94

	taboo scale	*thinking*	*extraversion*
female	3.67	3.27	
male	4.68	4.46	
indivudual administration	4.22		8.71
group administration	4.03		10.01
dichotomous	3.91		
six-categorical	4.28		

Table 3 gives the results based on the *taboo scale* and the remaining three MBTI scales (*Box*'s M-test: $F(110/\infty)=1.18$, $p=.101$). The level of significance was fixed at $\alpha=.05$.

Table 4 lists the F-statistics of univariate four-way layout analysis of variance in case of a significant effect. However, the p-values given can be interpreted only for reasons of description; they just give an impression of each scale's contribution to overall significance. Finally, if this contribution seems worthwhile, Table 5 shows the respective cell-means.

Table 6

Frequencies of agreement to three selected items of the *taboo scale* in dependence of sex and computer vs. paper-pencil administration (p-values according to *Pearson's Chi-square test*).

	item 5		item 6		item 9	
	SEX		SEX		SEX	
	f	m	f	m	f	m
computer	25	29	43	35	34	37
paper-pencil	19	31	20	37	13	35
	p= .511		p= .033		p= .037	

For detailed considerations some item-wise analyses have been done. Three selected taboo items with a medium rate of agreement (between 27% and 35%; that is item 5,6, and 9) turned out to depend on the interaction effect of CPxSEX as given in Table 6. According to *Pearson's Chi-square test* two of them confirm interaction effect.

DISCUSSION

While the pilot study rarely provides definite answers to our questions, the reported experiment does, indeed. Except for question 3), i.e., the question about the influence of the presence or absence of an investigator, all results are conclusive. Some of them are in fact surprising.

First of all, the answer to question 1) is „yes"; the main effect of individual vs. group administration is significant, and for this not only the *taboo scale* is responsible, but also the MBTI scale *extraversion vs. introversion.* As a matter of fact, testees surrounded with others exhibit not as much tabooed behavior as testees do at individual administration. And they seem to anticipate a higher social desirability for extraversive answers: the extraversion scores are higher at group than at individual administration. We think that this result alone destroys the concept of personality questionnaires.

Because of another significant main effect the answer to question 2) is „no". In addition to psychometric shortcomings which occurred in the pilot study the experimental results support the hypothesis of reactancy as a consequence of dichotomous item response format: at least with respect to the *taboo scale*, testees get at dichotomous response-formated items lower scores than at six-categorically response-formated items.

However, as one of the most interesting results the computer *per se* has no influence on the scores of a personality questionnaire (cf. question 4). Computer vs. paper-pencil administration is the only main effect that is not significant. That is, not even with very private and intimate items are there general differences in testees' answer behavior between computer and paper-pencil administration. The computer may, however, enhance or diminish effects that occur in any case: sex-based differences become less at the computer, and the six-categorical item response format becomes essential just then - and even that is not true for pertinent questionnaire-asked traits. Thus we abandon the idea that the computer does generally encourage a testee to give more valid answers.

There is no need to dwell on the third significant main effect, that is the more or less to be presupposed sex differences. But question 5) is to be answered with „yes" as already indicated. In regard to taboo items, female testees seem to answer to a computer more like male testees than they do with paper-pencil. A detailed interpretation of item-wise differences might, however, be artificial.

We think that very few moderator effects on a testee's answers to personality questionnaire items remain vague. According to the results of our pilot study the presence or absence of an investigator make perhaps any difference; or testees being from or not being from the target population differ; or, last but not least, testees being volunteers or being recruited differ. However, all that vagueness does not matter. The stated moderator effects disqualify personality questionnaires, indeed.

We therefore proclaim not to maintain personality questionnaires because of lack of alternatives. As recently Kubinger (1995) reviewed, this lack of alternatives is just a pretended one: There are various concepts of „objective personality *tests*" according to *R.B. Cattell.*

REFERENCES

BADER, P., HOFMANN, K., & KUBINGER, K.D. (1993). Zur Brauchbarkeit der Normen von Papier-Bleistift-Tests für die Computer-Vorgabe: Ein Experiment am Beispiel des Gießen-Tests [The fitness of standardizations of paper-pencil tests for computer administration: An experiment on the German-made Gießen-Test.] Zeitschrift für Differentielle und Diagnostische Psychologie, 14, 129-135.

BREHM, J.W. (1966). Theory of psychological reactance. New York: Academic Press.

EVAN, W.M., & MILLER, J.R. (1969). Differential effects on response bias of computer versus conventional administration of a social science questionnaire. Behavioral Science, 14, 216-227.

KOSON, D., KITTCHEN, C., KOCHEN, M., & STODOLSKY, D. (1970). Psychological testing by computers: Effects on response bias. Educational and Psychological Measurement, 30, 803-810.

KUBINGER, K.D. (1995). Objektive Diagnostik [Objective psychological testing]. In K. PAVLIK (Ed.), Enzyklopädie, Differentielle Psychologie: Grundlagen und Methoden [Encycplopedia, Differential Psychology: Foundations and methods] (pp. 507-541). Göttingen: Hogrefe.

KUBINGER, K.D. (1996). Zur Leichtgläubigkeit der Psychologen: Die unselige Anwendung von Persönlichkeitsfragebögen [Psychologists' credulity: The fatal use of personality questionnaires]. In M. Jirasko, J. Glück & B. Rollett (Eds.), Perspektiven psychologischer Forschung in Österreich [Psychological research work in Austria] (pp. 87-91). Wien: WUV-Universitätsverlag.

O'BRIAN, T., & DUDGALE, V. (1978). Questionnaire administration x computer. Journal of the Marked Research Society, 20, 228-237.

SCHWENKMEZGER, P., & HANK, P. (1993). Papier-Bleistift- versus computerunterstützte Darbietung von State-Trait-Fragebogen: eine Äquivalenzprüfung. [Paper-pencil vs. computerized testing of state-trait questionnaires: Proving the equivalency.] Diagnostica, 39, 189-210.

Received November 1999
Accepted December 1999