Review of Psychology
1996, Vol. 3, No. 1-2, 47-53 UDK 159.9

A new approach for the evaluation of convergent and discriminant validity

KLAUS D. KUBINGER

A novel procedure to evaluate convergent and discriminant validity of a test's or questionnaire's items is proposed. It works, in particular, for the case that the *Rasch*-model turns out not to conform the item-pool's apodictic content-validity. In addition, the procedure evaluates the effect of hypothesized moderator variables. All for this, non-parametric discriminant analysis must apply. The question is whether groups of subjects differing with respect to the given response to any item of the pool may be discriminated or not by their responses to the remaining items, and whether some dissimilar-construct variables or moderator variables do as well or even better. A given example serves for demonstration of the procedure. Though Kubinger's discriminant analysis was used in this paper, it is pointed out that the suggested approach is not restricted in this respect.

Although psychometrics yielded a fundamentally reorientation re the standards of psychological test-calibration in the last three decades, validation of a test is to claim the point of question, after all. Since *Rasch*'s and, perhaps, *Birnbaum*'s revolutionary discoveries of IRT (*item-response-theory*) the question of validity has almost been forgotten, at least depreciated. Of course, as for instance the *Rasch*-model must hold if a subject's score of solved items is to have any meaning (see e.g. Fischer, 1995), psychometric questions are of primary interest. However, even a *Rasch*-homogenous test means nothing unless is validity - it's whatever respect - is guaranteed. More than ever, a just claimed content-validity gains scepticism if a test does not fit the *Rasch*-model. Either just some of the items do not fit or the entire item-pool does not, even after some of the items have been deleted: While the first case reduces validity due to chance, the second case places doubt on the construct-validity of the item-pool as a whole.

This paper deals with these problems. It is based on Campbell and Fiske (1959) who combined, meritoriously, the concept of discriminant validity to the concept of convergent validity. The given paper is also based on some moderator effect concept. As a matter of fact, our approach is not based on the correlations of any test-*scores*, that is on the test in question as a fixed pool but is based on every singular item. As a matter of fact, the suggested procedure is not at all restricted to test calibrations according to the *Rasch*-model; the paper contributes to methods how to validate a test.

Klaus D. Kubinger, Institut für Psychologie, Univeristät Wien, Liebiggasse 5, A-1010 Wien, Austira (Correspondence concerning this article should be sent to this address).

Rasch-model-supported content-validity

Fortunately, the following method for developing a psychological test has already been established: First, items are to be generated according to some content-validity criteria, i.e. according to some authorities' ratings. Sometimes even certain generative rules are used, that is content-validity is implicitly defined by all combinations of material-components and cognitive operations, respectively. For instance, Formann and Piswanger (1979) offered an infinite pool of matrices test items by combining rules like *variating*, *superimposing*, and *sequencing* some symbols. Second, this (intuitively) stated content-validity is to be proven empirically by means of the *Rasch*-model. If the model holds, the items measure unidimensionally, indeed, they refer to a single trait. Keep in mind that, given a dichotomous response format, the *Rasch*-model is a sufficient condition for items measuring unidimensionally. - Keep also in mind that *Birnbaum*'s 2- and 3-PL model are two further probabilistic models which determine a unidimensional item-pool in case they hold. However, for both latter models no substantial model-check exists (cf. Kubinger, 1989). Therefore, they will not be taken into detailed consideration in this paper.

The way of developing a test, as described above, is conclusive and of tempting simplicity. Yet, the *Rasch*-model almost never holds initially; some items almost always have to be deleted. And - again almost always - chance ef-

fects are made responsible for model-contradictions. Hence, test authors being psychometricians usually try some kind of cross-validation by testing another sample with the reduced test and finally analyzing such data again according to the *Rasch*-model. However, admittedly, cross-validation does not always succeed. And even if it does, what is the flaw in those items which do not fit into the framework of the others, though content-validity originally seemed reasonable regarding them all?

Just Gittler (1986) demonstrated how the non-homogeneity of an item-pool may be explored: According to content analysis he tried an *a-posteriori* classification of the items of the German WISC subtest „Information" in order to re-analyze every subset of them by the *Rasch*-model - that is, after in another paper the item-pool as a whole has proven not at all to fit the *Rasch*-model. In this way the pool resulted so as to be partitioned into „knowledge of facts and book learning" and „everyday knowledge": When separated, each part suits the *Rasch*-model rather well, indeed!

Unfortunately, Gittler's approach can not be considered a panacea.

It should also be mentioned that Wottawa (1979) tried to pool items according to the *Rasch*-model by starting from pairwise homogeneous ones and adding one by one successively. However, this approach only delays the problems. So does a hierarchical analysis of dimensionality, that is, successive re-analyses of only those items which had to be deleted in the previous *Rasch*-model analysis. Finally, the concept of clustering the subjects before *Rasch*-model analysis take place in order to measure (different) traits group-specifically (cf. Rost, 1990), or the concept of evaluating any testee's appropriateness index in order to unmask its, the testee's instead of any item's contradiction to the *Rasch*-model (cf. Klauer, 1995), do not completely solve the problem in question.

That is, the *Rasch*-model is no method of validation, of course. At best, it gives colour to apodictically stated content-validity.

The basic idea of convergent and discriminant validity

According to Campbell and Fiske, a test not only needs a high affinity to other tests measuring the same or a similar construct but needs also a high diversity to such tests measuring some other construct. So far the test under consideration has to correlate with the former and, simultaneously, must not correlate with the latter. However, as just a single not fitting item diminishes (at least) convergent validity, their so-called „multitrait-multimethod matrix" means not really a proper starting point - this matrix consists of all intercorrelations of several traits or constructs, each measured by several (test-) methods. Neither are all revisions and generalizations of their original evaluation strategy proper. Ostendorf, Angleitner and Ruch (1986) give a review.

For instance, factor analysis fails to determine convergent and discriminant validity. No doubt, we expect that the test under consideration constitutes a common factor with all tests examining a similar construct; and we expect that all tests examining dissimilar constructs load different factors. However, the statistical problems with factor analysis are well-known and serious. There is no unequivocal criterion for the number of sufficient factors and there is no established statistical test which evaluates the loadings. And there is the problem with qualitative variables which might be of decisive importance - unless all variables are dichotomous (cf. Muthén & Christoffersson, 1981) qualitative variables cannot be included. That is, at least in case that interval-scaled, ordinal-scaled, and (multicategorical) qualitative data are intermixed, factor analysis does not work.

Probably, the unsatisfactory inventory of methods is the reason that the concept of convergent and, in particular, the concept of discriminant validity has not really taken hold. Of late, just the German edition of *Jackson*'s „Personality Research Form" (PRF; Stumpf, Angleitner, Wieck, Jackson & Beloch-Till, 1985) is deliberately based on this approach.

Since, on the other side, statistical methods are enriched by non-parametric discriminant analyses, convergent and discriminant validity should now be investigatable in a proper way.

Kubinger's non-parametric discriminant analysis

With reference to Kendall (1966), a non-parametric discriminant analysis was suggested by Kubinger (1983)[1]. This algorithm now handles interval- and ordinal-scaled as well as (multicategorical) qualitative data, simultaneously. Stepwise as well as overall procedure does work and significance tests apply with respect to any variable's contribution to discrimination. This test is based on a kind of cross-validation technique, that is, the percentage of correct group-allocations in one half of the sample is ascertained according to the allocation rules deduced from the other half. The algorithm has already stood the test in numerous empirical studies, though it is not very powerful because of its non-parametric conceptualization. Keep in mind that there are many alternative algorithms which, however, all refer to a less general data situation, but are consequently often more powerful. Maybe logistic regression (cf. Cox, 1970) as being implemented in SPSS is the

[1] A WINDOWS-version of the former FORTRAN-program by *Kubinger* is in preparation to be made available *via* internet by *Kubinger & Alexandrowicz.*

most attractive alternative. Of course, all of the following considerations also apply to any alternative.

The new approach

Let us now, actually, generalize the concept of convergent and the concept of discriminant validity to every item of a test or questionnaire.

Denote the k Items $X_1, X_2, \ldots X_j, \ldots X_k$; and suppose their realizations within a validation sample $i=1, 2, \ldots n$ to be $x_{ij}=0$ and $x_{ij}=1$, respectively. If all items measure one single and, above all, the same trait, then every one of them must be explained by the other. That is the meaning of convergent validity. If an item X_j is neither explained by one single item nor by all others, this item is not suitable within the framework of content-validity, contrary to contention. „Explaining" means discriminating the subjects with $x_{ij}=0$ and the subjects with $x_{ij}=1$ by realizations in X_l, $l=1, 2, \ldots (j-1),(j+1), \ldots k$.

Take into account some potential moderator variables $M_1, M_2, \ldots M_h, \ldots M_p$. That is, variables with an interaction effect on the test or some of its items, in particular with different interaction effects on different subgroups of the items. A typical example is gender as moderating certain intelligence-test items. Usually such moderator variables are ordinal-scaled or even (multicategorical) qualitative ones. Note that within the original „multitrait-multimethod matrix", moderator effects are neglected. If any M_h contributes significantly to discrimination with respect to $x_{ij}=0$ an $x_{ij}=1$ or even explains X_j better than all the other items of the test, it proves to be a moderator variable; otherwise it proves not to be.

Denote some, however-scaled, dissimilar-construct tests or items $Z_1, Z_2, \ldots Z_r, \ldots Z_s$. If X_j qualifies within the framework of claimed content-validity it must not be explained by any of these variables.

That is, we look for an arbitrary function $\boldsymbol{F}$ so that $X_j=\boldsymbol{F}(X_1, X_2, \ldots X_{(j-1)}, X_{(j+1)}, \ldots X_k; M_1, M_2, \ldots M_h, \ldots M_p; Z_1, Z_2, \ldots Z_r, \ldots Z_s)$. The question is, firstly, whether such a function exists, secondly, what type of variable effects mostly on X_j and, thirdly, what type of variable contributes at all. After answering these questions with respect to every item X_j, there might be a remaining pool of items which constitute a final test insofar as those items prove mutual convergent validity, while they do not depend on any hypothesized moderator variable or any dissimilar-construct variable. Remember, using *Kubinger*'s discriminant analysis, $\boldsymbol{F}$ prescribes some allocation rules. And keep in mind that a stepwise procedure like *Kubinger*'s is important for evaluating the multivariate hierarchical structure of dependencies.

An example

In a pilot-study, Kubinger (1978) developed a questionnaire for satisfaction of dwelling. Starting with 31 items, 19 finally proved to be *Rasch*-homogeneous, based on 166 subjects (see Figure 1). However, in the recently attempted cross-validation, based on 585 other subjects, unidimensionality could not be confirmed. There is not only a significant *Likelihood-Ratio-Test* when all 19 items are under consideration, but also when, in a second analysis, the worst three items have been deleted: $\chi^2=34.48$, $df=15$. As a result, the questionnaire does not fit the *Rasch*-model at the end. Its items are not compatible concerning the score attained, that is, the number of agreeing responses means nothing: Interpretation of agreement depends on the individual item presented.

The reason might be, on the one hand, that content-validity of the item-pool was originally not estimated by raters but just stated by the author. On the other hand, the reason for misfit might be based on the general problem of quantifying attitudes and personal attributes because of adulterations, that is at most according to social desirability. As a consequence of the latter reason *Rasch*-model analysis there does notseem to be worthwhile - keep in mind that adulterations almost never occur in achievement-tests.

1. I have no nice view from my residence.
2. The heating facilities in my residence are comfortable.
3. Inhabitants not belonging to the family-proper get on my nerves.
4. The electrical and hygienic installations in my dwelling are satisfying.
5. The flooring is as I like it.
6. There is some need of repair and renovation before our dwelling can come up to my standards for comfort.
7. The number and arrangement of the windows is well-done.
8. It bothers me that not all rooms are accessible from the hallway.
9. The arrangement of the equipment in the kitchen is well planned.
10. My residence offers no possibilities for personal privacy or pursuing hobbies.
11. There are enough walls free for cabinets and built-in furniture.
12. I would like a larger bath-room.
13. A connecting door between two certain rooms would certainly make life in this dwelling easier!
14. I'm satisfied by size and number of secondary rooms (cabinet, storage-room, cellar, and so forth).
15. I miss shower facilities.
16. It bothers me that not every family-member has his own separate room.
17. The arrangement of the rooms is not convenient (e.g. children's room not accessible from parents' bedroom, kitchen not from dining-room).
18. Some of our rooms are either too small or too big.
19. The legal status of our habitation (property, condominium, rental) satisfies me.

Figure 1: Questionnaire for satisfaction of dwelling by Kubinger (1978). For every statement subjects are asked either to agree or not.

also not being significant. Subjects with a high score or subjects who refuse to answer turn out to be proportionally more satisfied than the others. Suppose that a refuser refuses in order to hide his tendency towards „justification of expenditure" or does so in order to hide a habit of vacillation in opinion - the latter refuser would also not admit dissatisfaction as response to the questionnaire; then the items under discussion refer to a personality trait and fail to have discriminant validity.

Metaphorically speaking, the questionnaire vaporises. It has no meaning for any construct of satisfaction. As a measure of subjective quality of dwellings it is absolutely unfit. The hope that at least a subpool of items will prove to be valid cannot be confirmed.

DISCUSSION

Further research is needed before the cause can be identified: Are the destructive results specific to the example of dwelling satisfaction or does *Kubinger*'s discriminant analysis share the responsibility? The low power of this algorithm has already been stated. Furthermore, the algorithm comes close to assuming deterministic dependencies of the items, that is *Guttman*-scaled ones; so, even if the *Rasch*-model holds, items of medium „difficulties" would, perhaps, show no mutual convergent validity. However, keep in mind that any other pertinent discriminant analysis may be applied, as well. Shortcomings of the algorithm under discussion therefore do not necessarily mean that the entire approach undertaken in this paper is wholly inappropriate. On the other hand, this approach for evaluation convergent and discriminant validity will not apply to achievement tests very often because studies on this topic very seldom include a greater number of hypothesized moderator and dissimilar-construct variables - though such an application would settle the question of whether *Kubinger*'s algorithm stands the test or not. Initial research on the „3D-Cubes" of Gittler (1990) seems to speak for it. Most of the items prove there to be explainable.

The title of the paper suggests a *new* approach. Indeed, Campbell's and Fiske's idea was generalized from tests to test-items. But, of course, the approach remains extremely dependent on the actual sample like correlation and, in particular, factor analysis. And the well-known criticism of classical test theory mainly refers to correlations as to be the fundamental criteria of a test's quality. In this respect the given approach is not, strictly speaking, a new one, though, to be sure, a more elaborated one.

Concerning the criticism of correlations see also Künzel and Wottawa (1985). These authors point out the problem of traditional analyses not being able to distinguish between *correlative* relations, *necessary*, and *sufficient* conditions. Take in mind, that *Kubinger*'s algorithm is able to do so. And as Künzel and Wottawa use *Wottawa*'s HYPAG („hypotheses agglutination") in order to make an item-pool maximally supportive to some hypothesized path-model, their paper indicates that at least HYPAG is an alternative algorithm to *Kubinger*'s discriminant analysis.

REFERENCES

BECKMANN, D., BRÄHLER, E., & RICHTER, H. E. (1991). *Der Gießen-Test (GT).* [*The Gießen-Test.*] Berne: Huber.

CAMPBELL, D.T., & FISKE, D. W. (1959). Convergent and discriminant validity by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56*, 81-105.

COX, D. R. (1970). *The analysis of binary data.* London: Methuen.

FISCHER, G. H. (1995). Deviations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models - Foundations, Recent Developments, and Applications* (pp. 15-38). New York: Springer.

FORMANN, A. K., & PISWANGER, K. (1979). *Wiener Matrizentest (WMT).* [*The Viennese Matrices.*] Weinheim: Beltz.

GITTLER, G. (1986). Inhaltliche Aspekte bei der Itemselektion nach dem Modell von *Rasch.*[Rasch-modelba-sed item selection which happens by reasons of content.] *Zeitschrift für experimentelle und angewandte Psychologie, 33*, 386-412.

GITTLER, G. (1990). Dreidimensionaler Würfeltest (3DW). [3D-Cubes.] Weinheim: Beltz.

KENDALL, M. G. (1966). Discrimination and Classification. *Proceedings of an International Symposium on Multivariate Analysis.* New York: Akademic Press.

KLAUER, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models - Foundations, Recent Developments, and Applications* (pp. 97-110). New York: Springer.

KUBINGER, K. D. (1978). Der Einfluß der Planungsmitbestimmung auf die Wohnzufriedenheit. [The influence of participating in the planning phase on occupants' satisfaction with living in apartments.] *Psychologie und Praxis, 22*, 145-156.

KUBINGER, K. D. (1983). Some elaborations towards a standard procedure of distribution-free discriminant analysis. *Biometrical Journal, 25*, 765-774.

KUBINGER, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie. [A critical review of latent trait theory.] In K.D. Kubinger (Ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen.* [Modern test theory: A brief survey, with recent contributions.] Weinheim: Beltz, 19-83.

KÜNZEL, R., & WOTTAWA, H. (1985). Hinreichend, notwendig oder korrelativ? Bedingungen für das Zustandekommen von Leidensdruck und Therapiemotivation. [Sufficient, necessary, or associated? Conditions on suffering of pains and therapy motivation]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 6*, 175-184.

MUTHÉN, B., & CHRISTOFFERSSON, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407-419.

OSTENDORF, F., ANGLEITNER, A., & RUCH, W. (1986). *Die Multitrait-Multimethod Analyse.* [*The multitrait-multimethod analysis.*] Göttingen: Hogrefe.

ROST, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 3*, 271-282.

STUMPF, H., ANGLEITNER, A., WIECK, T., JACKSON, D. N., & BELOCH-TILL, H. (1985). *Deutsche Personality Research Form (PRF).* [German Personality Research Form.] Göttingen: Hogrefe.

WOTTAWA, H. (1979). *Grundlagen und Probleme von Dimensionen in der Psychologie.* [Fundamentals and problems of dimensions in psychology.] Meisenheim: Hein.

Accepted: December 1996