Review of Psychology
1998, Vol. 5, No. 1-2, 3-12 UDC 159.9

Psychometric shortcomings of Wechsler's intelligence scales - results on the German WISC, conclusions for the WISC-R

KLAUS D. KUBINGER

Based on psychometric models, especially the Rasch model, analyses of the German WISC and the German WISC-R show that hardly a single subtest scores fair. That is, the true extents of testees' abilities will not be represented correctly by the scores obtained under the given scoring rules. Since many of the items of the German WISC correspond to items of WISC-R, the same shortcomings must be suspected for this test battery. In this light, administration of these tests is no longer responsible. However, it is illustrated that Wechsler's basic concept is worthwhile when accompanied by (modern) psychometric tools: A new (German) test battery, AID, is introduced.

There is no doubt that Wechsler's scales represent a revolutionary progress in measuring intelligence: First of all, his definition of „IQ" (*intelligence quotient*) - refering to the general factor theory of intelligence - is still valid; secondly, his test batteries' polarization into *verbal* scales and *performance* scales convinces practitioners to this day. However, since the first issues of the Wechsler scales (beginning with Wechsler, 1939), psychometric standards have changed. Nowadays no psychological test may be applied or even published if the test does not clearly fit certain psychometric presuppositions. In 1939, on the other hand, an author's experience and intuition sufficed for apodictically fixing the concept and design of a test, the items, and the scoring rules.

In this paper only the scoring rules of Wechsler's scales are under discussion. The issue is fairness: do scores obtained indicate the true extent of a testee's ability? It must be made clear that scoring rules must not be surrendered to the test author's good faith: *they must be validated empirically*! If scoring rules are invalid, analyses of reliability, validity, discrimination and so on become merely of secondary interest, and hence are not under discussion here.

Because in the forties no psychometric model was available to a test author to check whether a given scoring rule is fair or not, the following criticism is not directed at David Wechsler personally. But at least since Georg Rasch's psychometric contributions in the sixties, no test author can be absolved from carelessness if he has not proved a scoring rule's fairness - for example Kubinger (1989) reviews applications of psychometric models for different cases. Hence, it is really a shame that, for instance, the revised children's scale of Wechsler (1974; the so-called WISC-R) does not benefit from such a check: the scoring rules of the subtests are as unproved and as apodictical as almost 40 years ago.

Concerning the German version of the former WISC, that is, the so-called HAWIK, empirical facts (Kubinger, Rop, Knoll & Wurst, 1983; better Kubinger, 1983) give impressive support to these doubts of the scoring rules' fairness: It is shown that hardly a single scoring rule of the eleven subtests proves fair. In consequence, this test battery may be said to measure nothing but nonsense. Steuer (1988) confirmed these results on another set of data and, moreover, got congruous results for the German version of WISC-R, that is, the so-called HAWIK-R.

As there is every reason to assume that, in the end, the same is true when considering WISC-R itself - of course, all adult scales, too - we sketch here the results obtained with HAWIK and HAWIK-R scales, respectively, and draw parallels to the WISC-R. The criticism is severe enough that the effort of analyzing this original test battery analogously does not seem worthwhile.

Data sets

The $HAWIK_1$ data set was collected from 1000 children, patients of a single child-guidance institution in Vienna. The children were aged between 6 and 15 years, almost two-thirds of them being male.

Data set $HAWIK_2$ was obtained from 936 children, patients of several child-guidance institutions in Vienna,

Klaus D. Kubinger, Institue of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria. (Correspondence concernig this article should be sent to this address).

Lower Austria, Berlin, Munich, and Erlangen (Germany). They were aged as well between 6 and 15 years; the percentages of male and female children were akin to the first data-set.

Data set *HAWIK-R* was obtained from 611 children, patients of institutions listed above and subjects of studies in Vienna, Hannover, and Landshut (Germany). Here the sexes were approximately the same in number, and the children were also aged between 6 and 15.

The following report focusses on the results pertaining to $HAWIK_I$ data.

The scoring rule of Information, Arithmetics, and Picture Completion

Of course, these subtests claim the testee's *score* to be an achievement-fair parameter. However, this implies that the number of solved items is a sufficient statistic, that is, irrespective of which of the subtest items have been solved (and which of them have not been solved) this number puts all testees' respective abilities in valid relations. Given so-called „local stochastic independence" - meaning that, there is no learning effect as one progresses through the test - then all subtest items should fit the Rasch model, the 1-PL model (Kubinger, 1989). As, indeed, within the subtests Information, Arithmetics, and Picture Completion no success-contingent learning effects take place, a Rasch model check will be a conclusive test of these subtests.[1]

Model checks were done as usual (Kubinger, 1989): *Andersen*'s Likelihood Ratio Test was used as a global check. Several cuts were made in the sample, most importantly the split between high scorers and low scorers. For visual materials the graphical model check was used, which served as well for (stepwise) item selection, in addition to *Fischer & Scheiblechner*'s z-test.

To summarize, none of the subtests in question turned out to fit the Rasch model.

The „best", but supportive, nevertheless crushing result concerns Picture Completion: This subtest contradicts the Rasch model only till half the items have been deleted.

In consideration of theories of intellectual development, re-analyses of both the other subtests were done with respect to age-homogeneous groups of testees. Again the scoring rule of Information proves not to fit. On the other hand, the Arithmetics' scoring rule does, indeed, fit the Rasch model within age-groups.

As a demonstration, consider the best result for the subtest Information in the $HAWIK_I$ sample: Restricted to the data of 10- and 11-year-old testees, a non-significant Likelihood Ratio Test turns out just if all but 13 items (out of 30) have been deleted (χ^2 (12) = 13.2, p> .05 „high vs. low score"). In contrast, when analyzing the entire sample, even a subtest reduced to nine items does not fit the Rasch model (χ^2 (8) = 49.5, p< .01; see Figure 1).

Based on the same data set ($HAWIK_I$), Gittler (1986) found in his re-analyses a tempting explanation of the lack of fit observed in the subtests Information and Picture Completion. For the first subtest, he defined intuitively two subgroups of items, i.e. „factual information or book learning" and „Information based on everday experience"; analyzing them separately according to the Rasch-model he succeeded in producing a fit - after slight corrections of item allotment. For the second subtest, he distinguished between items in wich missing part may be found by analogy, on the one hand, and items which require specific experiences for identifying the missing part, on the other hand. The latter succeeded in fitting the model after the deletion of just a few items.

However, Gittler's results do not imply that *Wechsler*'s original scales have to suit psychometric presuppositions. The opposite is true; that is, they reveal enormous failures

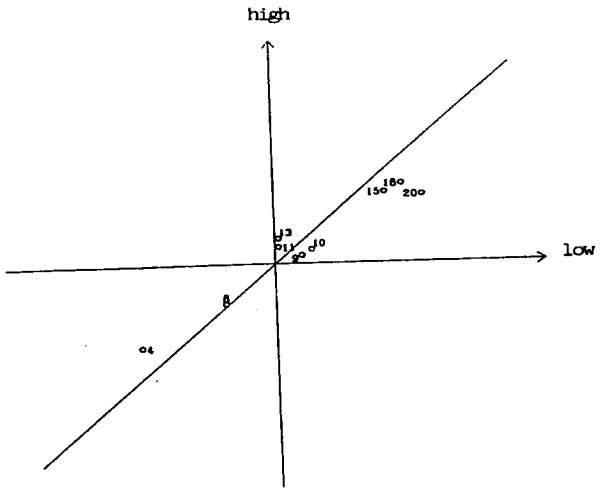

Figure 1. Graphical model-check: Item-parameter estimations of 9 Information-items, vertically based on testees with a high score, horizontally based on testees with a low score.

[1] In response to critical comments by a reviewer: Of course, the Rasch model only makes sense if the test in question is designed to measure a single dimension of ability. Indeed, very often practitioners claim a test measuring a single aptitude clearly isolated from others to be of practical unimportance. In all cases, though, the given scoring rule conditions all the deductions that can be made based on test results, whether unidimensionality is the target or just the opposite. If unidimensionality really is intended, the Rasch model is just one of many models which guarantees this; other examples are the 2- and 3-PL models (cf. for instance Kubinger, 1989). Absence of Rasch model fit demonstrates the invalidity of the given scoring rule only, but suspends the test that any other scoring rule

in item calibration. It might be of relevance to measure a testee's ability with respect to both the first and the second proposed definitions of „Information". However, these two almost independent dimensions of ability must not be mixed up: high ability in one will not compensate for low ability in the other.

Of course, even the existence of several items fitting the Rasch model in Picture Completion gives no support to *Wechsler*'s scales; the scoring rule as set forth in the manual is basically invalid.

Finally, aptness of the Arithmetics' scoring rule within age-groups does not pay: As a matter of fact within every age-homogeneous group only very few items discriminate. For instance, for 10- and 11-year-old testees, just six items showed solution probabilities not equal to zero or one. If the items are ranked according to their difficulty, great differences occur as compared to their presentation order (Table 1). As the test stands, a testee risks to be deprived of certain less difficult items which he could solve, because he fails in earlier but more difficult items.

18 HAWIK items from Information, 6 HAWIK items from Arithmetics, and 15 HAWIK items from Picture Completion correspond to WISC-R items and are thus of interest. Table II lists those which do not fit the Rasch model, that is, they score unfairly.

Table 2 gives rise to the feeling that some of Information's items are just miserably formulated or alternatively, the acceptable responses to them are too simplistic: for example, there is not really a „direction where the sun sets"; and „heat it" or „put it on the stove" is not really an explanation of what goes on when water boils.

Hence, the fact that older children or children with a high score find these items *relatively* more difficult than others is not surprising since they might be more thoughtful. Concerning Arithmetics, the trivial question on the number of pieces of a halved apple may similarily confuse the clever children. Concerning Picture Completion, in several items the missing part is not very difficult to *discover*, but its *name* is hard to find. Again younger children or children with a low score rely on *pointing* correctly while older ones or children with a high score might try to *name* the missing part but fail: „whiskers", „hinge", „center diamond", and „slot (slit, crack)"; that is, those items have to be deleted.

Table 1

Item difficulties according to the Rasch model as compared to their presentation ranking (10 items fitting the model with respect to 12- to 15-year-old testees; data set $HAWIK_1$).

Item number	Item parameter	(Rank of item parameter)
1		
2		
3		
4	-4.30	(1)
5		
6	-3.28	(2)
7		
8	-2.15	(3)
9	-1.64	(4)
10	1.88	(7)
11	-0.50	(5)
12	0.67	(6)
13	3.35	(10)
14	3.29	(9)
15	2.67	(8)
16		

Table 2.

WISC-R items of Information, Arithmetics, and Picture Completion for which the corresponding HAWIK items have proven to contradict psychometric presuppositions. For Information bear in mind that almost every item is contradictory if not analyzed age-wise.

Number Item	Psychometeric shortcomings
Information	
4 What must you do to make water boil?	8/9-years: high scorers find this item relatively more difficult than low scorers
7 How many days make a week ?	6/7-years: low scorers find this item incongruouslymore difficult than high scorers
11 What are the four seasons of the year ?	8/9-years: low scorers find this item incongruously more difficult than high scorers
14 In what direction does the sun set ?	10/11-years: Older ones find this item relatively more difficult than younger ones 12-15-years: high scorers find this item relatively more difficult than low scorers
20 How many pounds make a ton ? (German: How many kilograms make a ton)	8/9-years:low scorers find this item incongruously more difficult than high scorers
24 How tall is the average American man ?	12-15-years: high scorers find this item relatively more difficult than low scorers
Arithmetics	
5 If I cut an apple in half, how many pieces will I have?	High scorers find this item relatively more difficult the low scorers
13 A workman earned $ 36; he was paid $ 4 an hour. How many hours did he work ?	Low scorers find this item incongruously more difficult than high scorers
Picture Completion	
5 Cat	Older ones find this item relatively more difficult than younger ones
13 Door	High scorers find this item relatively more difficult than low scorers
14 Playing Card	High scorers find this item relatively more difficult than low scorers Older ones find this item relatively more difficult than younger ones
20 Screw	Girls find this item more difficult than boys Older ones find this item relatively more difficult than younger ones
25 Profile	Boys find this item more difficult than girls

Already singled out by the Rasch model, several items common to HAWIK and WISC-R Information, Arithmetics, and Picture Completion thus reveal content problems! We are thus confident that our strong reserves are valid for the WISC-R as well.

The scoring rules of Comprehension, Similarities and Vocabulary

If a testee's item responses are not simply categorized as „correct" or „incorrect", but partial credit scoring takes place, then psychometric presuppositions become even more elaborate. Hence the fairness of the Comprehension, Similarities, and Vocabulary apodictical scoring rules becomes even more unlikely. For these subtests every item response is scored on equidistant categories as either correct (2 points) or partially correct (1 point) or wrong (0 points).[2] Again, „local stochastical independence" is given; for this several psychometric models come into account (Kubinger, 1989).

The targeted analyses favoured Rasch's multicategorical *unidimensional* model, because as Andersen (1977) showed: If there exists a so-called „minimal sufficient statistic", that is, the estimation of the ability-parameter is based on only relevant and no irrelevant information and, in particular, the statistic does not ascertain item-wise testee's achievements, then this model must hold though restricted to a given scoring rule, an equidistant scoring. As a matter of fact, *unidimensional* models themselves presuppose that Rasch's multicategorical *multidimensional* model is valid: If this model is valid, indeed, the sum of item responses respective to any category would conclusively be a fair measure of a specific ability; „specific" means that this ability does not necessarily have to be the same across categories.[3]

Model checks were carried out by analogy.

We found that not even the *multidimensional* model proved to be valid regarding Comprehension and Similarities; and with regard to Vocabulary it proved valid only

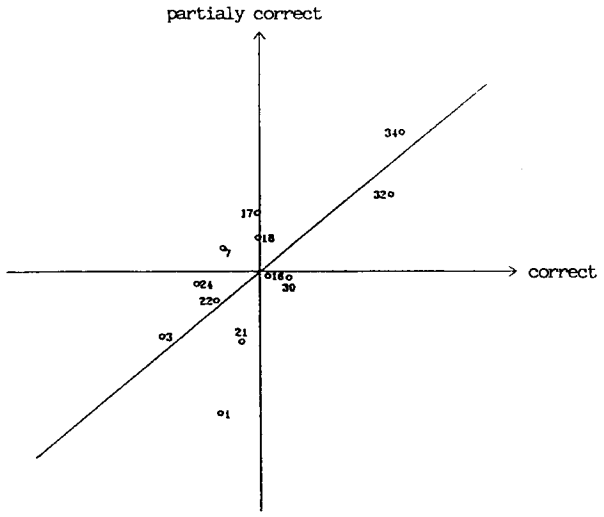

Figure 2. Graphical model-check: Parameter estimations of category „correct" (vertically) and category „partial correct" (horizontally); 12 Vocabulary-items.

afte r deleting at least 28 of 40 items ($\chi^2(22) = 26.0, p > .05$; „male vs. female" - $HAWIK_I$ data). However, the remaining items do not conform at all to the *unidimensional* model ($\chi^2(10) = 2554.1, p < .01$). Figure 2 shows that parameter estimations of the categories „correct" and category „partially correct" do not at all counterfeit the ratio 2:1! Moreover, it happens sometimes that a partially credited item response is harder to find than the corresponding correct one - as a result, the mean ratio (slope of the straight line) amounts to 1:0.88 instead of 1:0.5.

The scoring rules of the subtests under discussion are not fair; this conclusion stems not so much from inappropriate partial credit scoring but from items which are completely heterogeneous.

No Vocabulary item is common to HAWIK and WISC-R. On the other hand, 4 faulty items on Comprehension and 2 on Similarities are found in both batteries; these are listed in Table 3.

The scoring rules of Picture Arrangement, Block Design, and Object Assembly

These subtests provide also for partial credit scoring, this scoring differing however according to item. Moreover, the partial credit scoring does not always proceed equidistantly, in that it is based on more or less rapid (correct)

[2] Concerning HAWIK, this is true if we do not take the first four items of Similarities into account. Concerning HAWIK-R, this is true only for the subtest Similarities because the subtests Comprehension and Vocabulary contain dichotomously scored items - Steuer's (1988) analyses of Comprehension and Vocabulary, as described above by the Rasch model, did not lead to encouraging results either.

[3] Other authors, in particular Rost (1988), repel Rasch's multicategorical *unidimensional* model and prefer alternative models: Andrich's rating model, Masters' partial credit model, and, though not explicitely designated, Andersen's multicategorical model. The reason lies in loss of "specific objectivity": That is, in Rasch's model, estimation of the so-called category-parameter depends on the sample of testees used - whereas "specific objectivity" means that parameter estimation does not (statistically) depend on the chosen sample. However, using given category weights in *Rasch*'s model, here equidistant ones, according to Andersen (1977) invalidates this objection. At any rate, there is no need for a more detailed discussion because fit of Rasch's multicategorical *multidimensional* model is in any case a necessary condition.

Table 3

WISC-R items of Comprehension and Similarities for which the corresponding HAWIK items have proven to contradict psychometric presuppositions

Number	Item	Psychometeric shortcomings
Comprehension		
1	What is the thing to do when you cut your finger ?	Older ones find this item relatively more difficult than younger ones
6	What is the thing to do if a boy (girl) much smaller than yourself starts to fight with you ?	Older ones find this item relatively more difficult than younger ones Boys find this item more difficult than girls
7	In what ways is a house built of brick or stone better than one built of wood?	Younger ones find this item incongruously more difficult than older ones
9	Why are criminals locked up?	Younger ones find this item incongruously more difficult than older ones
Similarities		
6	In what way are beer and wine the same?	Low scorers find this item incongruously more difficult than high scorers
7	In what way are a cat and a mouse alike?	High scorers find this item relatively more difficult than low scorers

item responses.[4] For instance item responses of most of the items in HAWIK's Block Design are to be categorized as follows: wrong (0 points), correct but slow (4 points), and correct but quick (speed-graded 5, 6, or 7 points).

On the one hand, psychometricians have not yet thoroughly studied such a scoring rule; there is no proof that a certain model must hold necessarily for the scoring rule to be fair, nor that a certain model is sufficient for the scoring rule to be fair. On the other hand, because of the arbitrariness of the given category weights, there seems to be little hope for fairness.

At any rate, Andersen's multicategorical model (Kubinger, 1989) applies: As this model postulates that the category weights are given but these have not to be estimated, we might do some trial and error strategy in order to find those weights which probably come most closely to the adequate (fair) category weights. However, any such a result would not be conclusive.

Although originally Kubinger et al. (1983) as well as Steuer (1988) used an even more vague model for analyzing the three subtests under discussion, Huber (1986) used, indeed, Andersen's model as to *HAWIK-I* data and Block Design. The main result is: analysis failed because of the fact that within every partitition of the sample many solution probabilities were equal to zero or one. This was not the case for just 4 of 10 items, and for them quite surprising result was that actually Wechsler's category weights were those which suited the data best.

Nevertheless, there is no evidence that the scoring rules are fair. The given result indicates solely that any other category weights work even worse; but no finding exists that the respective subtests measure a certain unidimensional ability although the scoring rules intermix testee's „power" and „speed". Concerning WISC-R Block Design, scepticism even rises: This subtest consists of 9 HAWIK corresponding items but the weights differ from HAWIK; a correct but slow response scores just 3 points, which would

[4] As to HAWIK-R this is true only for Block Design and Picture Arrangement because item responses in Object Assembly are to be categorized uniformly and equidistantly with respect to three categories - however, analyses according to Rasch's multicategorical *multidimensional* model demonstrate unfairness of the scoring rule, as well.

be unfair if the analyses of Huber (1986) proved to be true. Maybe WISC-R Object Assembly have a fair scoring rule since partial credit scoring is based there not on bonus points for quickness but on the number of correctly joined cuts of the puzzle pieces.

The scoring rules of Digit Span and Coding

Both of these subtests have quite different scoring rules but we can treat them at once, because no analysis is possible and necessary, respectively.

For HAWIK both parts, Digits Forward and Digits Backward, are to be scored according to the largest number of correctly repeated digit series - with no respect to the sum of trials needed. As this number represents the limit of memory capacity it is obviously a fair mean to compare the achievement of different testees. Yet, this fact does not pay if such numbers were summed up for Digit Forward and Digit Backward. That is, in doing so the same psychometric presuppositions arise as in the preceding chapter, though just two „items" (with 8 categories each) are not at all analyzable.

Within WISC-R as well as within HAWIK-R the testee has a second trial even if a certain number of a digit series is correctly repeated the first time; and the total sum of correctly passed trials in Digit Foreward and Digit Backward makes up the score. Of course, it is not self-evident that repeating a certain length of a digit series both times means twice as much memory capacity as repeating it just one time. Thus Andrich's binomial model (Kubinger, 1989) applies here; if it holds, this scoring rule would be fair - concerning HAWIK-R, Steuer (1988) unfortunately missed such a check.

The scoring rule of Coding needs no psychometric analysis because the number of coded symbols means nothing else than the intensity of speed and is therefore fair. All to demand are enough items, because the bonus points provided for early completion are again critical, that is, their appropriateness and fairness, are not proved but stated apodictically only. As a matter of fact, *HAWIK_I* data showed that more than 6 percent of 6- and 7-aged testees profit from bonus points in Coding A, hence this subtest scores perhaps unfairly, too.

Since Mazes of WISC-R were not transfered to HAWIK and HAWIK-R, we give no comment on this subtest.

An example of a test-battery fulfilling psychometric presuppositions: AID

The (German-made) test-battery AID (Adaptive Intelligence Diagnosticum; Kubinger & Wurst, 1985) demonstrates that it is possible to prosecute Wechsler's subtest ideas and still conform to psychometric presuppositions. The items of all respective subtests have been calibrated according to psychometric models which guarantee the scoring rule in question to be fair, given the model empirically actually holds.

To be sure, though the subtests of AID were based on Wechsler subtest, the latter were to undergo a thorough and general content re-conceptualization. But this will not be considered here.

Indeed, the subtests corresponding to Information, Comprehension, Arithmetics, Similarities, Vocabulary, Picture Completion, Picture Arrangement, and Block Design resulted as to fit the Rasch model. Likelihood-Ratio-Test and graphical model-check for instance for the subtest Applied Mathematics illustrates that thorough item creation does pay: Starting with 80 items, 63 items fitted the model at the end ($\chi^2(62)= 84.4$, $p<.05$; „high vs. low score"; see Figure 3). And the Object Assembly corresponding test proved to conform Rasch's multicategorical *multidimensional* as well as Rasch's multicategorical

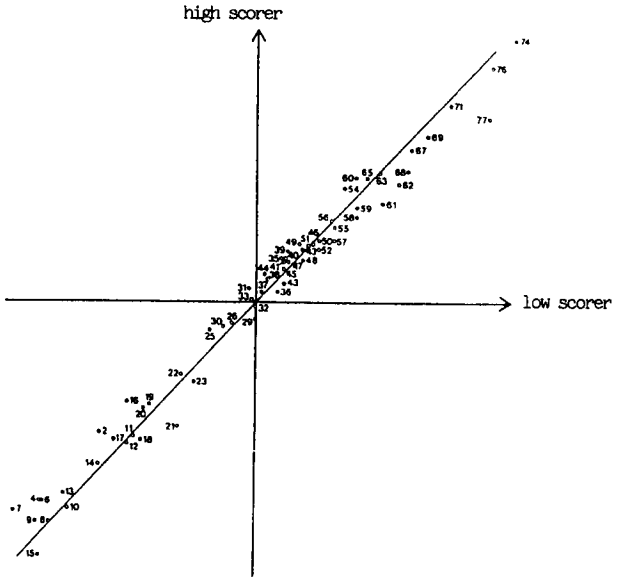

Figure 3. Graphical model-check: Item-parameter estimations of 63 items of AID-subtest „applied mathematics", vertically based on testees with a high score, horizontally based on testees with a low score.

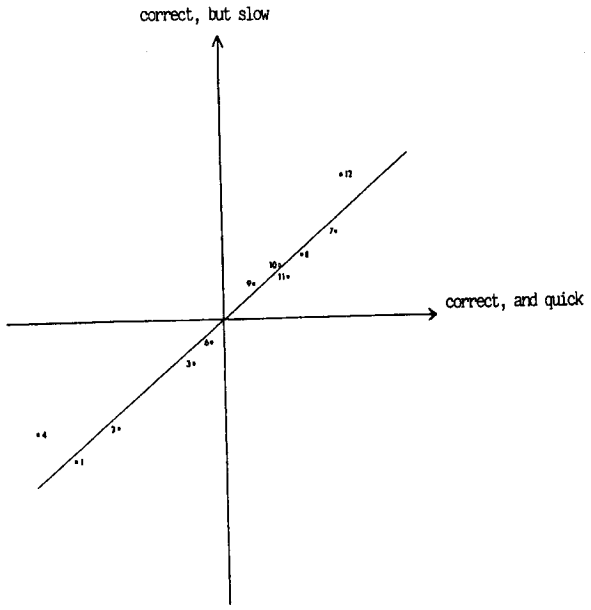

Figure 4. Graphical model-check: Parameter estimations of category „correct, but slowly" (vertically) and category „correct, but quick" (horizontally); 11 items of AID-subtest corresponding to Object Assembly.

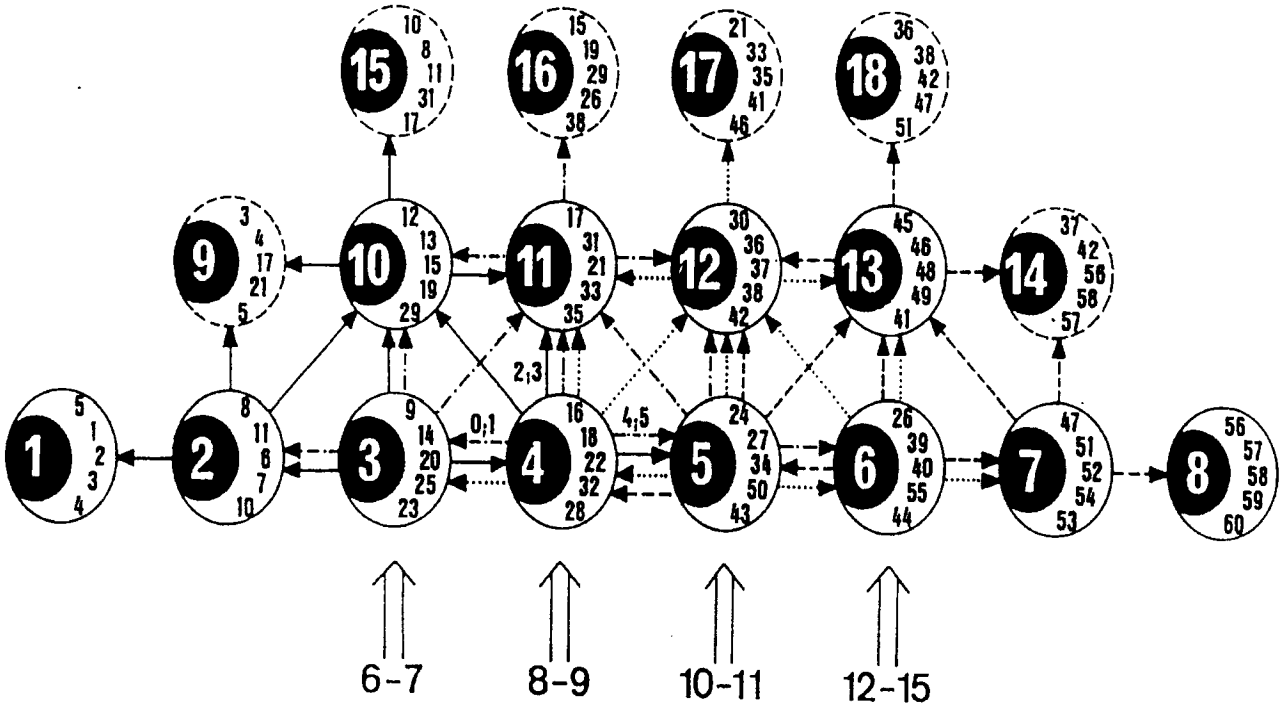

Figure 5. The branched testing design for AID. Circles represent different subsets up-leveled from left to right. Each subset contains five items, the number of each corresponding to the rank of difficulty. The age of the testee determines the starting point. As branched testing terminates after the third subset, dashed-line subsets consists of items of solid-line subsets.

unidimensional model, the latter indeed with equidistant category weights (see Figure 4) although surprisingly, power and speed measure here a common ability.

With regard to Digit Span and Coding the scoring rules of HAWIK are retained in AID apart from using two separate scores for „forward" and „backward" in addition, the number of items to code is large enough.

Moreover, AID realizes adaptive testing. With regard to HAWIK we criticized above that many items have no discriminatnt power within homogeneous groups of testees; hence, administration which concentrates item testing at the individual level of a testee's ability is to be preferred. Meanwhile, principles of adaptive testing (e.g. Kubinger, 1987) are very familiar so that just the idea of so-called branched-testing is sketched below (for more details concerning AID see Kubinger, 1988). Starting with an age-conformed first subset of 5, items a second followed by a third subset of 5 items is to be administered, the latter depending on the preceding score of the testee (Figure 5). Because all items suit the Rasch model, it is possible to estimate the asked-for ability parameter (which allows fair comparisons of all testees), even if they performed on completely different items. And of course, standard error of measurement (i.e. estimation) is for branched testing much smaller than for testing conventionally. Five subtests of AID work in this way, four more in a similar but simplified branched-design. The subtests corresponding Digit Span and Coding do not work adaptively.

DISCUSSION

To practitioners it seems hard to welcome the given results and general objections. In particular, HAWIK(-R) fans and the more WISC-R fans tend to reject any objection at least for reasons of justifying their efforts. Surely, it is hard to bear psychometric facts while, with regard to contents, the tests work convincingly well. However, be sure HAWIK and HAWIK-R score certainly unfair and WISC-R is very implausible in not doing so!

By the way, the most crucial scoring rule, that is, the definition of IQ, was not yet touched. Since Wechsler stated the IQ as the unweighted average of the subtests' scaled scores, any multicategorical *unidimensional* (Rasch-„related") model must hold. The results of a forthcoming paper based on $HAWIK_I$ data contradict, however.

As we belong to the fans of Wechsler's basic concept, this paper was not at all intended to smash this concept, although its realizations, especially HAWIK(-R), should be smashed. We tried to argue that using and elaborating this concept but taking deliberately psychometric prerequisites into account leads to a proper test-battery. And, indeed, AID already stood the test.

REFERENCES

ANDERSEN, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69-81.

FISCHER, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* [*Introduction into psychomterics.*] Bern: Huber.

GITTLER, G. (1986). Inhaltliche Aspekte bei der Itemselektion nach dem Modell von Rasch. [Rasch model-based item selection and the matter of content.] *Zeitschrift für Experimentelle und Angewandte Psychologie, 33*, 386-412.

HUBER, M. (1986). *The general latent structure model for contingency table data: Simulationsstudien und Untersuchung des Mosaik-Tests des Hamburg-Wechsler Intelligenztests für Kinder.* [*The general latent structure model for contingency table data: Analysis of Block Design of the Germnan WISC.*]. Unpublished doctoral dissertation, University of Vienna, Vienna.

KUBINGER, K.D. (1983). Konstruktive Kritik am HAWIK. Ausgangspunkt für das Konzept eines neuen Tests. [Criticism of the German WISC: The very beginning of a new test.] *Zeitschrift für Differentielle und Diagnostische Psychologie, 4*, 203-221.

KUBINGER, K.D. (1987). Adaptives Testen. [Adaptive testing.] In R. Horn, K. Ingenkamp & R.S. Jäger (Eds.), *Tests und Trends 6* (pp. 103-127). München: PVU.

KUBINGER, K.D. (1988). On a Rasch-model-based test for noncomputerized adaptive testing. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 277-289). New York: Plenum.

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

KUBINGER, K.D., ROP, I., KNOLL, E. & WURST, E. (1983). Ergebnisse der testtheoretischen Analyse des HAWIK. [Psychometric analysis of the German WISC.] In K.D. Kubinger (Ed.), *Der HAWIK - Möglichkeiten und Grenzen seiner Anwendung.* [*The*

German WISC - its practicability (pp. 115-186)]. Weinheim: Beltz.

KUBINGER, K.D. & WURST, E. (1985). *Adaptives Intelligenz Diagnostikum (AID).* [*Adaptive intelligence diagnosticum.*] Weinheim: Beltz.

ROST, J. (1988). *Quantitative und qualitative probabilistische Testtheorie.* [*Quantitative and qualitative analysis by latent trait theory.*] Bern: Huber.

STEUER, O. (1988). *HAWIK und HAWIK-R: Testtheoretische Analysen des HAWIK und seiner revidierten Form als Wiederholungsstudie und Weiterführung der Arbeit von Kubinger (1983), „Der HAWIK - Möglichkeiten und Grenzen seiner Anwendung".* [*German WISC and WISC-R: psychometric analysis according to Kubinger (1983) „The German WISC - its practicability.*] Unpublished doctoral dissertation, University of Vienna, Vienna.

WECHSLER, D. (1939). *The measurement of adult intelligence.* Baltimore: Williams & Wilkins.

WECHSLER, D. (1974). *Manual for the Wechsler Intelligence Scale for Children- Revised.* New York: Psychological Corporation.

Received July 1998
Accepted October 1998