Psychopathology *17*: 49–58 (1984)

0254–4962/84/0172–0049$2.75/0

Development of a New Shortened Version of Raven's Matrices Test for Application and Rough Assessment of Present Intellectual Capacity within Psychopathological Investigation

R. Wytek, E. Opgenoorth, O. Presslich

Psychiatrische Universitätsklinik Wien, Österreich

Abstract. Based on a sample of 300 psychiatric patients the items of the Standard Progressive Matrices test are analyzed in terms of classical and probabilistic methods, and a version shortened to 30 items is developed. This new version of the test is then standardized from a new sample of 1,200 patients. A table of selected percentiles is computed. Validation with respect to rough classification of intelligence is proved by comparison with results of the WIP.

The extent to which a patient is capable of handling tests and questionnaires independently frequently poses a problem in the clinical application of psychological diagnostic methods, for their validity depends upon satisfactory execution. Not only must the patient possess a certain amount of intelligence in order to handle most of the objective personality tests, he must also be capable of sustaining a sufficient level of attention throughout independent test tasks, and be able to work without relying on personal feedback from the tester.

For reasons of economy in psychological outpatient work, a testing method meeting such conditions should take a minimal amount of time, be minimally verbal so that patients with little formal education can be explored as well, and above all be able to differentiate sufficiently in the lower scattering range of the norm.

The experience of the Psychological Laboratory of the Vienna University Psychiatric Clinic shows that the scope of Raven's standard progressive-matrices test (1958) would meet these requirements, assuming that a new standardization of this test were available. In comparison to other tests, for example the CFT3, it is not so demanding on the patient, so that those patients with low standard or substandard intelligence can handle the test without undue difficulty. Observation of the patients' behavior during test execution permits further conclusions to be drawn concerning their capacity for work as well as the interpretability of other tests they may have accomplished. Experience with the use of this test shows further, that the first

Table I. Item analysis

Results of the classical test analysis				Results of the probabilistic analysis of successive item selection, model test					
original item	probability of item solution	item standard deviation	index of selectivity	1	2	3	4	5	6[1]
01	0.990	0.099	0.074	*					
02	0.983	0.128	0.220	*					
03	0.967	0.180	0.262	*					
04	0.960	0.196	0.310	*					
05	0.957	0.204	0.294	*					
06	0.973	0.161	0.305	*					
07	0.863	0.343	0.437	*					
08	0.863	0.343	0.399	*					
09	0.937	0.244	0.397	*					
10	0.870	0.336	0.443			*			
11	0.757	0.429	0.501				*		
12	0.513	0.500	0.544						
13	0.983	0.128	0.190	*					
14	0.930	0.255	0.402	*					
15	0.927	0.261	0.383	*					
16	0.843	0.363	0.366		*				
17	0.830	0.376	0.525						
18	0.777	0.416	0.484						
19	0.673	0.469	0.525				*		
20	0.503	0.500	0.691		*				
21	0.550	0.497	0.672						
22	0.600	0.490	0.725		*				
23	0.473	0.499	0.700		*				
24	0.347	0.476	0.550						
25	0.903	0.296	0.457	*					
26	0.867	0.340	0.443						
27	0.800	0.400	0.526						
28	0.663	0.473	0.640						
29	0.740	0.439	0.695						
30	0.663	0.473	0.606						
31	0.683	0.465	0.601						
32	0.587	0.492	0.638						
33	0.497	0.500	0.540						
34	0.430	0.495	0.594						
35	0.263	0.440	0.388		*				
36	0.157	0.363	0.348		*				
37	0.883	0.321	0.484	*					
38	0.707	0.455	0.688						
39	0.690	0.462	0.707						
40	0.643	0.479	0.639						
41	0.767	0.423	0.657						
42	0.640	0.480	0.726		*				
43	0.560	0.496	0.696						
44	0.563	0.496	0.601						
45	0.567	0.496	0.651						
46	0.433	0.496	0.631						
47	0.293	0.455	0.427		*				
48	0.120	0.325	0.362						

Table I (cont.)

Results of the classical test analysis				Results of the probabilistic analysis of successive item selection, model test					
original item	probability of item solution	item standard deviation	index of selectivity	1	2	3	4	5	6[1]
49	0.530	0.499	0.622						
50	0.443	0.497	0.641						
51	0.440	0.496	0.674						
52	0.323	0.468	0.594						
53	0.360	0.480	0.610						
54	0.340	0.474	0.503					*	
55	0.343	0.475	0.342		*				
56	0.303	0.460	0.528						
57	0.207	0.405	0.494						
58	0.087	0.281	0.296					*	
59	0.037	0.188	0.087		*				
60	0.047	0.211	0.074		*				

[1] No further items selected.

series of the test (items 1–12) offers apparently no information because practically all patients can solve it. Therefore we proceeded first of all with an item analysis of the test, with the purpose of shortening it at the same time.

Intended Purpose of the Investigation

Our experience with this test has made it obvious to us that a reduction in the number of items still permits a satisfactory rough classification of intelligence, on condition that it not be used to measure high and very high intelligence.

Although the first items of the Raven test can usually be correctly solved, its total of 60 items requires a relatively long solving time. Tests which require a lot of time and a great amount of concentration are usually too exacting for the patient, making an accurate estimation of his intellectual capacity rather subsidiary. A shortening of the test is thus justified not only from the standpoint of the therapist but also on the grounds of test methodology. Raw data indicate that item reduction should have practically no bearing on the test's validity, because very easy tasks occur side by side with very difficult ones, and test theoretical expectations are not fulfilled.

A shortening of the test is thus the first objective. The second, the establishment of selected percentiles, is of equal importance. For this purpose we feel that a rough classification is not only sufficient, but also adequate for the test. It should be possible to approximate the results to the 'IQ', thus conforming to the usual measures of intelligence and facilitating communication with the doc-

tor. The percentiles arrived at should represent a rough estimation for the applicability of other paper-pencil tests, as well as the information contained therein of the patient's capacity to concentrate, his ability to work independently, and naturally also his intelligence.

The standardization of the new version is based on a patient population of our psychiatric clinic, the statistics gathered embracing not only the clientele of the University Psychiatric Outpatient Clinic, but also 'testable' inpatients.

For the standardization of the new shortened form of the Raven test, data from healthy probands was *not* collected, because in our opinion the usefulness of a test for clinical purposes cannot be evaluated in reference to a general population. The relative position of the individual patient within a sufficiently big patient population permits, in our opinion, sufficient conclusions as to his intellectual capacity to be drawn. This hypothesis should be controlled by means of a cross validation with the help of the WIP according to *Dahl* [1972].

Owing to the selection of the sample, the results of the investigation are only applicable to a clinical population; nevertheless, certain results should be generalizable, for example, the results from the Rasch analysis, because Rasch's model distinguishes itself as the only one of test theory that is independent of sampling. The application of the probabilistic model of Rasch [*Fischer,* 1974], and the classical test theory [*Lienert,* 1969], above all in reference to certain item characteristics, are thus accepted on methodical grounds. The final selection of the items should follow from the results of classical test theory as well as those of the Rasch analysis.

Method

First Step: Test Shortening

The statistical and test theoretical calculations were carried out on two different samples. The first sample of 300 patients served for the test shortening and Rasch tabulation. Although the item choice taken from classical test theory methods arrived at very similar results compared to the successive adap-

Table II. Rasch conform items

Original item	Probability of solution	Index of selectivity	Rank of item according to solution probability
12	0.513	0.535	19
17	0.830	0.479	2
18	0.777	0.461	4
21	0.550	0.652	17
24	0.347	0.551	26
26	0.867	0.410	1
27	0.800	0.527	3
28	0.663	0.638	11
29	0.740	0.703	6
30	0.663	0.633	10
31	0.683	0.584	9
32	0.587	0.656	13
33	0.497	0.536	20
34	0.430	0.592	24
38	0.707	0.686	7
39	0.690	0.701	8
40	0.643	0.638	12
41	0.767	0.638	5
43	0.560	0.695	16
44	0.563	0.611	15
45	0.567	0.655	14
46	0.433	0.640	23
48	0.120	0.365	30
49	0.530	0.625	18
50	0.443	0.648	21
51	0.440	0.688	22
52	0.323	0.615	27
53	0.360	0.632	25
56	0.303	0.542	28
57	0.207	0.507	29

tation of the test on the Rasch model, the results obtained on the basis of the Rasch model were given absolute priority, and immediately integrated into the clinical routine. This was done owing to their particular advantages, namely probabilistical test model, sample independence in the estimation of item parameters, specific objectivity of the Rasch tabulated tests, and above all the possibility of model tests as opposed to all earlier employed classical methods.

Table I shows the classical characteristic values for the 60 items of the original test, as well as the determination of the test shortening and adaptation to the Rasch model carried out in six steps. Lower and higher test values were used as criteria for the model adaptation.

Table II gives characteristics of the Rasch-conform items.

Table III shows the change of reliability in the course of successive item selections.

Second Step: Percentile Distribution

The test, shortened to 30 items, was applied to a sample of 1,200 unselected patients in order to put the percentile distribution from the first sample (n = 300) on a broader basis.

5 percentile levels were defined in order to categorize test performance (table IV).

As performance in intelligence tests depends on the age, 5 age categories were defined. In order to estimate the score distribution in the 5 age categories, and judge their accuracy, the 'bootstrap' method by *Efron* [1979] was employed. This method is basically similar to the well-known jackknife methods for evaluating the variance of statistics.

Table V shows the results obtained by the bootstrap method – estimation of the mean and its variation for the 5 age categories and for the percentile level defined in table IV.

Table VI shows the age-dependent norm distribution for the new shortened form of the Raven test.

Table III. Mean, standard deviation, and reliabilities according to the item analysis and item selection

Total number	Number of selected items	Mean	Standard deviation	Split-half reliability
60	–	36.65	13.08	0.967
46	14	23.53	11.98	0.963
35	11	19.33	9.67	0.960
34	1	18.64	9.53	0.960
32	2	17.03	9.07	0.953
30	2	16.60	8.73	0.956

Table IV. Percentile levels

Intelligence	Raw distribution cumulative percentages	Percentile levels
Very low	5	1
Low	25	2
Average	75	3
High	95	4
Very high	100	5

Table V. Means and their variance (±SD) as a result of the bootstrap method for the percentile levels and age categories

Age, years	Percentile levels			
	1	2	3	4
15–24	9.17 ± 2.72	20.07 ± 1.28	27.55 ± 0.47	29.35 ± 0.50
25–34	7.34 ± 2.10	17.69 ± 0.85	26.18 ± 0.54	28.91 ± 0.37
35–44	3.95 ± 1.30	13.85 ± 1.14	24.04 ± 0.73	28.27 ± 0.49
45–54	3.40 ± 1.21	10.63 ± 1.23	23.04 ± 0.85	27.51 ± 0.78
55–	1.95 ± 1.07	9.28 ± 2.35	20.88 ± 1.23	25.52 ± 1.11

Third Step: Rough Classification of Intelligence

In order to elucidate to what extent the pragmatically defined 5 percentile levels enable conclusions as to performance in a frequently employed intelligence test to be drawn, the data of every patient, with whom testing both with the new Raven short form and with the WIP (short form of the Hamburg Wechsler Intelligence Test by *Dahl*) was carried out, was selected from the routinely stored data of the Psychological Laboratory of our Clinic.

A priori IQ expectations for the individual percentile levels of the Raven short form were defined as follows:

Percentile niveaus	IQ
1	–80
2	81–90
3	91–110
4	111–120
5	121–

The WIP-IQ values for the patients who actually took this test were transformed in a 5-part scale corresponding to the one given above; the results were compared to the new form of the Raven test.

In tables VII–IX contingency charts are presented for evaluating the comparability between the results of the WIP and the Raven short form. The total sample of n = 390 was compiled from patients belonging to the original standardization sample and patients belonging to the 'new' standardization sample (both WIP-IQ and Raven short form results were gathered for 144 patients from the first standardization sample embracing 300 patients, and for 246 patients from the second one embracing 1,200).

Discussion

The first pragmatically defined percentile boundaries (5, 25, 75 and 95%) for the new version of the Raven test are defined in such a way that a broad average group, which embraces 50% of the patient population, is demarcated from two marginal groups on either side. The marginal groups embrace 20 and 5%, respectively, of the sample. The estimation of the means in table V shows that a markedly smaller variation of test performance exists in the upper regions of intelligence, that is, that the new form of the test discriminates less in this level of intelligence than in lower levels of intelligence. This corresponds to one of the purposes we had in mind.

It is also notable that age dependency is much more pronounced in the lower intelligence level than in the higher within the same percentile boundaries. The precision of the

Table VI. Norm distribution (n = 1,200 unselected patients)

Age, years	Number of patients	Percentile levels				
		1	2	3	4	5
15–24	255	0–9[1]	10–20	21–28	29	30
25–34	324	0–7	8–18	19–26	27–29	30
35–44	344	0–4	5–14	15–24	25–28	29–30
45–54	162	0–3	4–11	12–23	24–28	29–30
55–	115	0–2	3–9	10–21	22–26	27–30

The upper interval limit coresponds to the mean value estimation in the bootstrap method.

[1] Number of solved items.

Table VII. Table of contingency between percentile levels derived from results of Raven matrices and WIP (taken from the total sample n = 390. Partial sample: standardization sample n_1 = 144; 'new' standardization sample n_2 = 246)

Raven test (30 items)	WIP											
	rank 1		rank 2		rank 3		rank 4		rank 5		Σ	
Rank 1												
15–24 years	3		4		2						9	
25–34 years	6		7		2						15	
35–44 years	3	19	3	17	5	13					11	49
45–54 years	3		2		1						6	
55– years	4		1		3						8	
Rank 2												
15–24 years	3		11		14		3				31	
25–34 years	3		5		11		1				20	
35–44 years	4	15	4	29	9	61		5	1	3	18	113
45–54 years	4		4		12		1		1		22	
55– years	1		5		15				1		22	
Rank 3												
15–24 years			5		35		23		4		67	
25–34 years			2		14		6				22	
35–44 years	1	3	2	13	18	90	8	48	1	14	30	168
45–54 years	1		3		16		6		3		29	
55– years	1		1		7		5		6		20	
Rank 4												
15–24 years					1		5				6	
25–34 years					8		6		3		17	
35–44 years				3	5	18	1	18	2	11	8	50
45–54 years			1		3		5		3		12	
55– years			2		1		1		3		7	
Rank 5												
15–24 years					3		2		1		6	
25–34 years							1				1	
35–44 years						3		3	1	4	1	10
45–54 years									2		2	
55– years											0	
Σ												
15–24 years	6		20		55		33		5		119	
25–34 years	9		14		35		14		3		75	
35–44 years	8	37	9	62	37	185	9	74	5	32	68	390
45–54 years	8		10		32		12		9		71	
55– years	6		9		26		6		10		57	

Statistic by *Kullback and Leibler* [1951], transformed into χ^2 = 166.8 (significant).

Table VIII. Table of contingency between percentile levels derived from results of Raven matrices and WIP (partial sample: $n_1 = 144$)

Raven test (30 items)	WIP											
	rank 1		rank 2		rank 3		rank 4		rank 5		Σ	
Rank 1												
15–24 years	2		3		1						6	
25–34 years	2		2								4	
35–44 years	2	8		7	2	3					4	18
45–54 years			1								1	
55– years	2		1								3	
Rank 2												
15–24 years			7		7		2				16	
25–34 years	1		2		3						6	
35–54 years	2	7		11	1	18		2	1	3	4	41
45–54 years	4		2		4				1		11	
55– years					3				1		4	
Rank 3												
15–24 years			2		20		10		1		33	
25–34 years			1		6		3				10	
35–44 years				3	9	41	2	17		3	11	64
45–54 years					4		1		1		6	
55– years					2		1		1		4	
Rank 4												
15–24 years							2				2	
25–34 years					2		4		1		7	
35–44 years				1	2	4		9		1	2	15
45–54 years							3				3	
55– years			1								1	
Rank 5												
15–24 years					3		1		1		5	
25–34 years							1				1	
35–44 years						3		2		1	0	6
45–54 years											0	
55– years											0	
Σ												
15–24 years	2		12		31		15		2		62	
25–34 years	3		5		11		8		1		28	
35–44 years	4	15	0	22	14	69	2	30	1	8	21	144
45–54 years	4		3		8		4		2		21	
55– years	2		2		5		1		2		12	

$\chi^2 = 77.3$ (significant).

Table IX. Table of contingency between percentile levels derived from results of Raven matrices and WIP (partial sample: $n_2 = 246$)

Raven test (30 items)	WIP											
	rank 1		rank 2		rank 3		rank 4		rank 5		Σ	
Rank 1												
15–24 years	1		1		1						3	
25–34 years	4		5		2						11	
35–44 years	1	11	3	10	3	10					7	31
45–54 years	3		1		1						5	
55– years	2				3						5	
Rank 2												
15–24 years	3		4		7		1				15	
25–34 years	2		3		8		1				14	
35–44 years	2	8	4	18	8	43		3			14	72
45–54 years			2		8		1				11	
55– years	1		5		12						18	
Rank 3												
15–24 years			3		15		13		3		34	
25–34 years			1		8		3				12	
35–44 years	1	3	2	10	9	49	6	31	1	11	19	104
45–54 years	1		3		12		5		2		23	
55– years	1		1		5		4		5		16	
Rank 4												
15–24 years					1		3				4	
25–34 years					6		2		2		10	
35–44 years				2	3	14	1	9	2	10	6	35
45–54 years			1		3		2		3		9	
55– years			1		1		1		3		6	
Rank 5												
15–24 years							1				1	
25–34 years											0	
35–44 years								1	1	3	1	4
45–54 years									2		2	
55– years											0	
Σ												
15–24 years	4		8		24		18		3		57	
25–34 years	6		9		24		6		2		47	
35–44 years	4	22	9	40	23	116	7	44	4	24	47	246
45–54 years	4		7		24		8		7		50	
55– years	4		7		21		5		8		45	

$\chi^2 = 108.2$ (significant).

mean value evaluations for both marginal groups of the age distribution (15–24 and above 55 years) is less in the lower level of intelligence (SD = 2.72 and 2.35, respectively).

On the one hand this can be plausibly explained through the selection of the patient population of our clinic, namely, few young patients whose intellectual endowment is extremely poor. On the other hand the higher distribution concerning poorly endowed older patients cannot be explained using the same argument, and should therefore be attributed to an effect of sampling. Nevertheless, the distribution in both cases is small enough so that no difficulties could arise on practical application. The contingency charts for the investigation of the compatibility between Raven and WIP results (results from both tests transformed in rank values) show statistically a very high agreement between both tests (table VII).

About 41% of the sample (namely 160 from 390 patients) have identical ranks in both tests (main diagonal), a further 48% (186 patients) have test results whose difference is limited to one rank. 89% of the sample is thus accounted for. When one takes the result of the WIP intelligence classification as a standard for the concept of 'general intelligence', the new Raven short form underestimates the 'WIP intelligence' three times more often than it overestimates it (172 underestimations, 58 overestimations).

An explanation for this distinction may be that the WIP also picks up verbal, that is premorbid or crystallized intelligence, and the Raven test picks up intelligence of less established nature, more easily disturbed through psychopathological changes. This interpretation is supported by the comparison of the underestimations and overestimations in both contingency charts for test data stemming from driver's licence examinations testing former psychiatric patients whose behavior was 'strange', and for such test data concerning psychiatric patients with current psychic disturbances (table VIII–IX).

Results for 'driver's licence patients' give a proportion of 54 underestimations to 20 overestimations, a proportion corresponding to 2.7:1, and for the other psychiatric patients a proportion of 118 to 38, respectively, 3.1:1. Nevertheless, the agreement between WIP and Raven for these two partial samples is practically as good as for the total sample (table VII).

When one uses the total raw values as opposed to our purpose of creating a rough classification of intelligence, a product-moment correlation of the original Raven total raw values with the WIP-IQ receives a value of 0.61.

For the new short form of the Raven progressive-matrices test a product-moment correlation with the WIP-IQ results in a value of 0.59. Reduction of the test items from 60 to 30 is thus justifiable in this respect as well.

References

Dahl, G.: WIP. Reduzierter Wechsler-Intelligenztest (Hain, Meisenheim am Glan 1972).

Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Statist. *1:* 1–26 (1979).

Fischer, G.: Einführung in die Theorie psychologischer Tests (Huber, Bern 1974).

Kullback, S.; Leibler, R.A.: On information and sufficiency. Ann. math. Statist. *22:* 79–86 (1951).

Lienert, G.A.: Testaufbau und Testanalyse (Beltz, Weinheim 1969).

Raven, J.C.: Standard progressive matrices (Lewis, London 1958).

O. Presslich,
Psychiatrische Universitätsklinik Wien,
Währinger Gürtel 76–78,
A–1090 Wien (Austria)