British Journal of Clinical Psychology (1995), **34,** 237-250 *Printed in Great Britain*

The WAIS-R(UK): Basic psychometric properties in an adult UK sample

J. R. Crawford*

School of Psychology, Flinders University of South Australia, GPO Box 2100, Adelaide SA5001, Australia

C. D. Gray and K. M. Allan

Department of Psychology, University of Aberdeen, Scotland

The WAIS-R is the most widely used measure of intellectual ability in the UK, despite never having been standardized in this country. The present study examined the psychometric properties of the WAIS-R in a sample of 200 subjects, which was representative of the adult UK population in terms of the distributions of age, sex and social class. The properties of the three IQ scales, i.e. the FSIQ, the VIQ and the PIQ, were found to be very similar to those reported for the US standardization sample: the scores were normally distributed, with means close to the desired value of 100; moreover, the reliabilities of the IQ scales were extremely high and closely matched the US reliabilities. There were also indications, however, that the scales have restricted standard deviations when used in the UK. The reliabilities of the 11 original subtests ranged from moderate to high and the majority were similar to the US reliabilities. However, in addition to evidence of restricted SDs, significant differences (sometimes as much as two-thirds of an SD) were found among the subtest means. These in-built subtest discrepancies could lead to erroneous conclusions about an individual's performance. A conversion table for UK test users is provided to overcome this problem.

The Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) has now been superseded by the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981). Although the WAIS-R retains the basic format of the WAIS, the rules for the administration and scoring of some subtests have been modified and a substantial proportion of the item content has been updated. The latest version of the Wechsler scales was standardized on a highly representative sample of 1880 Americans during the period 1976-80 (Wechsler, 1981). The WAIS-R, like its predecessor, has come to be regarded as a core test in clinical practice. As Kaufman notes, 'The WAIS—R is *the* criterion of adult intelligence, and no other instrument even comes close' (Kaufman, 1983; p. 313).

In the UK, the WAIS-R is the test most widely used by clinical psychologists. A recent illustration of the importance of the WAIS—R in decision making is provided

* Requests for reprints.

by the report from The British Psychological Society (BPS) on 'mental impairment' (BPS, 1991). This report arose in response to the need for operational criteria for 'mental impairment' and 'severe mental impairment', following legislative changes (to the Mental Health Act). The report recommends that the WAIS-R be used as the *sole* measure for identifying the presence of mental impairment.

The present study was prompted by two facts: firstly, the WAIS-R has not been standardized in the UK; secondly, despite this, and as noted, the test is widely used by UK psychologists. A UK supplement to the WAIS—R test manual is now available (Lea, 1986); but this simply incorporates replacements for, or modifications to, unsuitable US test items (e.g. dollars and cents are replaced by pounds and pence in the Arithmetic subtest). In the absence of a UK standardization, UK psychologists, when interpreting WAIS—R scores, must assume that the psychometric properties of the test in the UK population are the same as those of the US standardization sample. This is clearly an unsatisfactory state of affairs, since the number of currently untested assumptions is considerable. For example, the aforementioned BPS report proposed that a Full Scale IQ (FSIQ) score that is more than two standard deviations (SDs) below the mean should be the criterion for 'mental impairment'. The report also stressed that confidence intervals should accompany the reporting of IQ scores, and that particular attention should be given to these in borderline cases. These recommendations rest upon the following assumptions: that WAIS—R scores are normally distributed in the UK; that the US standardization has fortuitously set the UK population mean and SD at the desired values of 100 and 15, respectively; and that the US and UK reliabilities of the test are also equivalent (the last assumption is implicit, because the construction of a confidence interval around an individual IQ score relies upon the value of the *standard error of measurement* (SEM), which, in turn, is a function of the test's reliability and SD).

Most clinicians in the UK, when interpreting an individual's WAIS—R performance, seek not only to compare the FSIQ score with the population mean but also to examine the *profile* of scores, that is, to identify the individual's relative strengths and weaknesses. Such analysis involves a further set of assumptions. For example, if the psychologist wants to determine whether there are statistically significant differences among the subtest scores, or between the Verbal (VIQ) and Performance (PIQ) scales, a crucial assumption is that the *population* means on these components are equal. For the US psychologist, given the quality of the standardization sample, this assumption is eminently reasonable, since the WAIS—R was standardized to have these very characteristics. However, we currently have no information with which to assess the correctness of this assumption in the UK. Thus, in the UK, discrepancies among an individual's subtest scores cannot be taken as evidence of underlying differential abilities, since such discrepancies may be artifactual, in the sense that they may simply reflect differences among the subtest *population* means in the UK.

The aim of the present study was to examine the validity of the general assumption that the WAIS—R, when used in the UK, has the same psychometric properties as it has in the USA. The assumption was tested by recruiting a sample of 200 healthy subjects, representative of the UK population in terms of age, sex and social class. The specific questions posed were as follows:

(1) Are the distributions of the three WAIS—R scales normally distributed in the UK? Additionally, is the distribution of VIQ—PIQ discrepancies also normal?

(2) Are the means and variances for the three IQ scales equal to one another and to the corresponding values for the US standardization sample? Is this also true for the 11 subtests?

(3) Do the IQ scales and the subtests have adequate reliabilities and are these equal to the corresponding values in the USA?

Method

Subjects

Two hundred subjects, (104 female, 96 male), screened by interview for the presence of neurological or psychiatric disorder, participated in the present study. All subjects were resident in the North-East of Scotland, and most were urban dwellers. Most received a small honorarium for their participation. Subjects were recruited from a wide variety of sources, e.g. local and national businesses, clubs (pensioners' clubs, angling clubs), community centres, etc.

The mean age of the sample was 44.3 years (SD = 19.2 years), with a range from 16 to 83 years. The mean number of years of education was 12.6 years (SD = 3 years), with a range from 7 to 21 years. (In calculating the number of years of education, subjects were credited with half a year for each year spent in part-time education.) The social class of each subject was derived from their occupation, using the Classification of Occupation (OPCS, 1980). Retired subjects were coded according to their previous occupations. Unemployed subjects were coded by their previous occupation, or as Social Class 5 if they had never worked. Subjects describing themselves as housewives/househusbands were coded by their previous occupation, or by their spouse's occupation if they had never worked.

The recruitment strategy was intended to obtain a sample that was broadly representative of the adult population in the UK, with respect to the distributions of social class, age and sex. To determine the extent to which this aim had been met, the social class distribution in the present sample (Table 1) was compared with that of the UK adult population, according to the 1981 Census.

Table 1. Social class distribution in the WAIS–R sample and in the adult UK population (percentages)

	Social class				
	1	2	3	4	5
WAIS–R sample	7	27	42	17	8
General adult UK population	5	23	48	18	6

A chi-square goodness-of-fit test showed that the class distribution in the present sample did not differ significantly from the population distribution (^^{2}(4) = 5.66, n.s.).

A similar procedure was adopted to examine the representativeness of the sample in terms of age distribution. Nine age bands were formed, corresponding to those adopted for the WAIS—R standardization sample, with the exception that the 70—74 age band was replaced with a 70 + band. The percentage of subjects in each band are presented in Table 2, along with Census-derived expected percentages. A goodness-of-fit test revealed that the observed and expected distributions did not differ significantly (,\;^{2}(4) = 7.71, n.s.).

Finally, a goodness-of-fit test showed that the sex distribution in the sample did not differ significantly from the Census proportions (^^{2}(1) = 0.01, n.s.).

All subjects were administered a full-length UK WAIS-R, in accordance with standard procedure (Lea, 1986; Wechsler, 1981). Raw scores were converted to scaled scores which, in turn, were summed to derive IQs. As was noted above, examination of an individual's profile of scores at the subtest level is generally viewed as being as important as (or indeed more important than) examination of scores on

Table 2. Age distribution in the WAIS–R sample and in the adult UK population (percentages)

	Age distribution								
	16–17	18–19	20–24	25–34	35–44	45–54	55–64	65–69	70+
WAIS–R sample	3	3.5	11	22.5	14.5	12	11.5	8	14
General adult UK population	4.5	4.3	9.7	18.7	15.8	14.8	14.61	6.5	11.3

the IQ scales. When conducting such an examination, it is inappropriate to use the normal scaled scores, as those are not corrected for the effects of age (the correction for age is incorporated when the sum of scaled scores is converted to an IQ). Instead, scores should be converted to *age-graded* scaled scores, using the tables provided in the WAIS-R manual (Wechsler, 1981, p. 35). Subtest age-graded scaled scores were calculated for the present sample, and the analysis of subtest means and SDs were based on these, rather than upon the normal scaled scores.

Kolmogorov-Smirnov goodness-of-fit tests (Siegel, 1956, pp. 47-52) were used to determine whether the empirical cumulative distributions of scores for the FSIQ, VIQ and PIQ deviated significantly from the appropriate normal cumulative distribution. The same procedure was used with the empirical cumulative distribution of VIQ—PIQ discrepancies.

In the present study, the reliabilities of all the WAIS—R subtests (with the exception of Digit Symbol) were estimated using the split-half method: that is, the correlations between odd and even items were computed and the Spearman—Brown Prophecy Formula applied to correct for halving the size of the item sample (Anastasi, 1988, p. 121). The split-half method was also used by Wechsler (1981) for all the subtests except Digit Symbol and Digit Span.

Because of the nature of the Digit Symbol test, it was not possible to use the split-half method to estimate its reliability. Instead, as in Wechsler (1981), a randomly selected subsample was retested and the reliability estimated by the test—retest method. In the present study, the size of the subsample was 46 and the median test-retest interval was 63 days. The mean age of the subsample was 46.9 years (SD = 19.4 years). The mean number of years of education was 12.3 years (SD = 2.6 years).

The reliabilities of the individual subtests having been established, Mosier's (1943) formula for the reliability of a composite was used to estimate the reliabilities of the FSIQ, VIQ and PIQ. A more readily accessible presentation of the formula, which was used by Wechsler (1981) to estimate the equivalent reliabilities for the US standardization sample, can be found in Nunnally (1978, p. 248).

Estimates of the SEMs for each subtest and the three IQ scales were computed from their sample SDs and estimated reliability coefficients, according to the usual formula (Anastasi, 1988, p. 133).

Results

Tests for normality of the distributions of FSIQ, VIQ and PIQ

The results of the Kolmogorov—Smirnov tests are presented in Table 3 *a.* It can be seen that none of the empirical distributions of scores on the FSIQ, VIQ or PIQ differs significantly from the normal distribution; indeed, the very high*p*-values lend additional support to the interpretation that the three population distributions are close to normal.

The outcomes of the Kolmogorov—Smirnov tests are consistent with the bell-shaped appearance of graphs and displays of the three distributions and the values of other statistics of the distributions of FSIQ, VIQ and PIQ. From Table *3b,* it can be seen that the means and medians have very similar values in all three distributions. Furthermore, the statistics of skewness and kurtosis give no reason to reject the hypothesis of normality.

Table 3. *(a)* Comparison of the cumulative distributions of the three samples of IQ scores (FSIQ, VIQ and PIQ) with the cumulative normal curve N(100, 255): results of the Kolmogorov—Smirnov goodness-of-fit test

IQ	Most extreme difference	z	p
FSIQ	0.059	0.83	.50
VIQ	0.047	0.66	.77
PIQ	0.036	0.51	.96

(b) Means, medians and measures of skewness and kurtosis

			Skewness		Kurtosis	
	Mean	Median	Value	Value/SE	Value	Value/SE
FSIQ	102.45	102.00	0.20	1.138	0.04	0.108
VIQ	102.44	103.00	0.14	0.809	−0.39	−1.130
PIQ	102.03	102.00	0.09	0.504	−0.11	−0.313

Mean scores on the IQ scales and subtests

The mean scores on the IQ scales and the 11 subtests are presented in Table 4. The desired values for these means are, of course, 100 and 10, respectively. In the present sample, the mean FSIQ is 102.5 (SD = 13.1). The 95 per cent confidence interval is (100.18, 103.82), which does not include the value 100. The 99 per cent confidence interval, however, is (99.59, 104.41). There are grounds for viewing the value 102.5 as an overestimate of the UK population mean. It will be noted that (despite the nonsignificance of the chi-square value) there is a slight over-representation of Social Classes 1 and 2 in the UK sample (see Table 1). In view of the well-corroborated correlation between IQ and social class (in the present sample the Pearson correlation is .60), marginal inflation of the estimate of the population value is to be expected.

Turning now to comparison of the VIQ and PIQ scales, a related-samples / test showed that the mean scores on the two scales did not differ significantly *(t(* 99) = 0.54, p = .59). Moreover, this is a very powerful test, as can be seen from the narrowness of the 99 per cent confidence interval, which is (—1.57, 2.39). For all practical purposes, therefore, it can be assumed that, in the UK, the VIQ and PIQ scales have the same mean.

Theoretically, the normality of two distributions implies that normality of any linear function of those distributions, and so, having accepted the hypotheses of normality for the VIQ and PIQ distributions, one would expect the distribution of the VIQ-PIQ discrepancies to be normal also. Acceptance of the null hypothesis, however, does not prove that it is true: the Kolmogorov—Smirnov tests of VIQ and PIQ considered separately may have failed to pick up departures from normality which may summate to render the distribution of VIQ-PIQ discrepancies non-

Table 4. Mean scores, SDs and results of chi-square tests comparing the sample variances with US population variances

IQ scale	Mean	SD	χ^2	Cumulative probability
(*a*) The three IQ scales				
Full Scale IQ	102.45	13.12	152.24	.0058
Verbal IQ	102.43	12.81	145.05	.0015
Performance IQ	102.03	13.39	158.47	.0156
(*b*) The eleven subtests				
Information	10.08	2.99	197.19	.4771
Digit Span	11.36	2.80	172.93	.0909
Vocabulary	10.39	2.38	125.71	.0000
Arithmetic	11.31	2.65	155.64	.0101
Comprehension	10.16	2.66	156.47	.0115
Similarities	9.80	2.48	135.78	.0002
Picture Completion	10.07	2.85	179.10	.1588
Picture Arrangement	10.78	2.95	192.99	.3932
Block Design	10.94	2.75	167.38	.0501
Object Assembly	9.58	2.55	144.28	.0013
Digit Symbol	10.59	2.34	120.91	.0000

Table 5. Results of a within-subjects ANOVA of the data from the 11 WAIS–R subtests

Source	d.f.	SS	MS	*F*	p^a
Subtest	10	697.33818	69.73882	16.62	< .00005
Subjects	199	7383.35439			
Error	1990	8351.93455	4.19695		
Total	2199	16432.62712			

[a] This *p*-value is the conservative Greenhouse–Geisser value.
Bartlett's sphericity statistic = 159.50, with 45 d.f., $p < .0005$.
F_{max} criterion = 2.71, with (10, 199) d.f.
F_{max} obtained = 1.63.

normal. Furthermore, it has been considered necessary to examine empirically whether the distribution of VIQ—PIQ discrepancies in the US standardization sample conforms to the normal curve (Matarazzo & Herman, 1984,1985). In the present UK sample, therefore, the Kolmogorov-Smirnov test was also applied to the distribution of VIQ—PIQ differences. The outcome of the test indicates that the distribution of discrepancies is normal (most extreme difference = 0.05; ^ = 0.71; *p* = .70). The indications are, therefore, that, in the UK population, VIQ-PIQ discrepancies are normally distributed, with a mean of zero.

To determine whether there were significant differences among the subtest means, a within-subjects analysis of variance (ANOVA) was carried out upon the 11 correlated subtest samples. The ANOVA summary is given in Table 5.

It will be noted that, while the hypothesis of homogeneity of variance must be accepted on Hartley's F_{max} test, Bartlett's test shows that the hypothesis of sphericity must be rejected, with a small*p*-value. This necessitates a more conservative test, the most conservative of which is the Greenhouse—Geisser adjustment of the degrees of freedom. Even on the Greenhouse—Geisser test, however, the null hypothesis of equality of the subtest means is rejected, with a very small />-value.

Following a repeated measures ANOVA, multiple paired comparisons among the individual treatment means require appropriate protection against undue inflation of the per family Type I error rate. In view of the non-sphericity of the data (as shown by the Bartlett test), there is a risk of inflation of the per family Type I error rate if, as in the Tukey HSD test (Tukey, 1953), a single error term is used for all comparisons (Jaccard, Becker & Wood, 1984). In such circumstances, Jaccard *et al.* recommend a Bonferroni procedure described by Myers (1979, p. 303), in which a value / is calculated for each pair of treatment conditions, *using only the scores in the two conditions concerned* (as in the ordinary, paired-samples *t* test), and setting the critical value at ^$_{2a}$/a(a-i)> where *a* is the number of treatment conditions in the experiment.

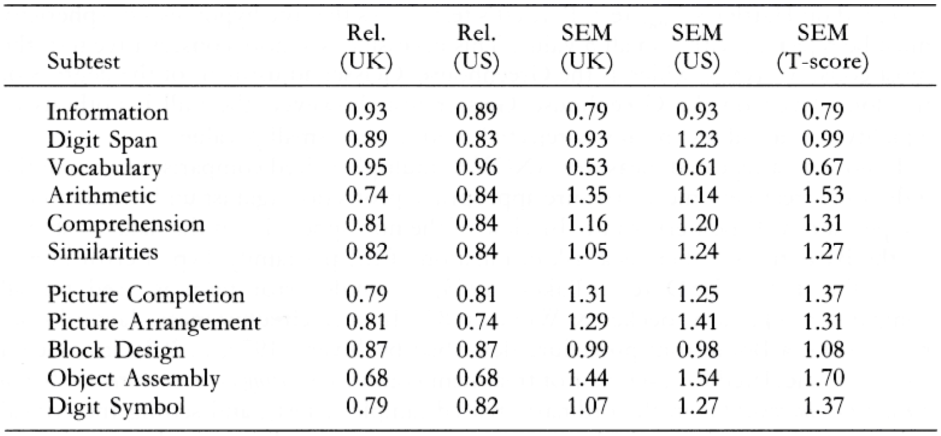
Table 6 shows the values of / for the differences between all the 55 possible different pairs of means that can be drawn from the entire array of 11 subtest means. It can be seen that 24 of the 55 pairs of means differ significantly and, in some cases (e.g. Object Assembly paired with Arithmetic), the mean difference amounts to almost two scaled score points (i.e. around two-thirds of an SD). Thus, the indications are that, in the UK, the WAIS—R subtest means are not equivalent to each other.

Table 6. The (absolute) values of *t* for multiple pairwise comparisons among the 11 subtest means

		OA	Sim	PC	I	Comp	V	DSy	PA	BD	Ar	DSp
Subtests	Means	9.58	9.80	10.07	10.08	10.16	10.39	10.59	10.78	10.94	11.31	11.36
OA	9.58	—	1.10	2.41	2.23	2.59	*3.95*[a]	*4.75*	*5.40*	*7.61*	*7.68*	*6.89*
Sim	9.80		—	1.60	1.61	2.25	*4.28*	*4.01*	*5.33*	*6.05*	*7.82*	*7.57*
PC	10.07			—	0.03	0.40	1.69	2.36	*3.52*	*4.22*	*5.39*	*5.24*
I	10.08				—	0.43	2.03	2.23	3.48	*4.09*	*6.23*	*5.26*
C	10.16					—	1.72	2.12	2.87	*3.77*	*5.80*	*5.63*
V	10.39						—	1.10	2.12	2.86	*4.91*	*5.03*
DSy	10.59							—	0.79	1.68	3.38	3.36
PA	10.78								—	0.78	2.50	2.44
BD	10.94									—	1.68	1.80
Ar	11.31										—	0.26
DSp	11.36											—

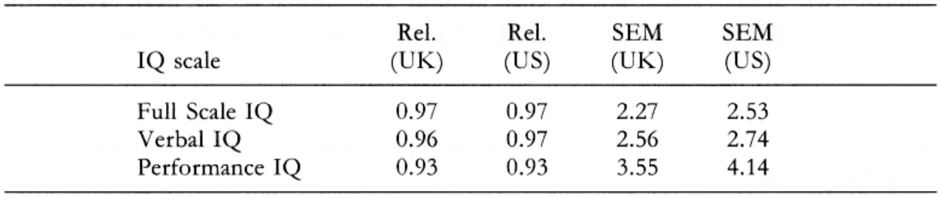
[a] The italicized entries all exceed 3.5, the critical value of *t* for Myers–Bonferroni unplanned paired *t* tests (with per family Type I error rate set at 0.05).

Table 7. Reliabilities and SEMs for the 11 subtests in the WAIS–R UK sample. For the purposes of comparison, the reliabilities and SEMs for the US sample are also presented



Subtest	Rel. (UK)	Rel. (US)	SEM (UK)	SEM (US)	SEM (T-score)
Information	0.93	0.89	0.79	0.93	0.79
Digit Span	0.89	0.83	0.93	1.23	0.99
Vocabulary	0.95	0.96	0.53	0.61	0.67
Arithmetic	0.74	0.84	1.35	1.14	1.53
Comprehension	0.81	0.84	1.16	1.20	1.31
Similarities	0.82	0.84	1.05	1.24	1.27
Picture Completion	0.79	0.81	1.31	1.25	1.37
Picture Arrangement	0.81	0.74	1.29	1.41	1.31
Block Design	0.87	0.87	0.99	0.98	1.08
Object Assembly	0.68	0.68	1.44	1.54	1.70
Digit Symbol	0.79	0.82	1.07	1.27	1.37

Table 8. Reliabilities and SEMs for the three IQ scales in the WAIS–R UK sample. For the purposes of comparison, the reliabilities and SEMs for the US sample are also presented



IQ scale	Rel. (UK)	Rel. (US)	SEM (UK)	SEM (US)
Full Scale IQ	0.97	0.97	2.27	2.53
Verbal IQ	0.96	0.97	2.56	2.74
Performance IQ	0.93	0.93	3.55	4.14

SDs/variances of the IQ scales and subtexts

The SDs of the IQ scales and subtests are presented in Table 4. To determine whether the sample estimates of the UK SDs differed significantly from the US values, the SDs were converted to variances and a chi-square test performed on the ratio between the population and sample variances, multiplied by the degrees of freedom (Howell, 1987). The chi-square values and their probabilities are also presented in Table 4. It can be seen that all three IQ scales yield significant chi-square values. Thus, the null hypothesis that the samples were drawn from a population with a variance of 225 (i.e. SD = 15) must be rejected. It would appear, then, that the UK population SDs for the IQ scales are significantly smaller than the desired values. The same procedure was followed for the 11 subtests, and the results are also presented in Table 4. It can be seen that, for the majority of subtests, the results indicate that the UK population variances are less than the desired value of 9 (i.e. SD = 3).

An *F* test was performed to determine whether there was a significant difference

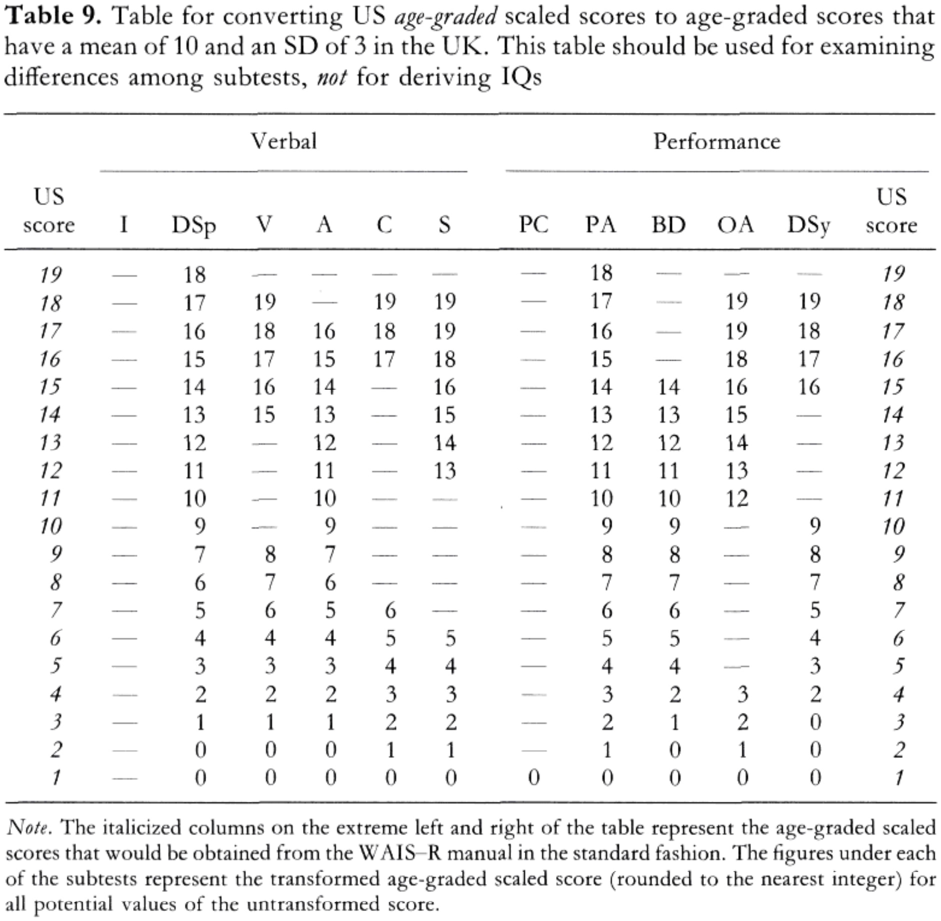


Table 9. Table for converting US *age-graded* scaled scores to age-graded scores that have a mean of 10 and an SD of 3 in the UK. This table should be used for examining differences among subtests, *not* for deriving IQs

	Verbal						Performance					
US score	I	DSp	V	A	C	S	PC	PA	BD	OA	DSy	US score
19	—	18	—	—	—	—	—	18	—	—	—	*19*
18	—	17	19	—	19	19	—	17	—	19	19	*18*
17	—	16	18	16	18	19	—	16	—	19	18	*17*
16	—	15	17	15	17	18	—	15	—	18	17	*16*
15	—	14	16	14	—	16	—	14	14	16	16	*15*
14	—	13	15	13	—	15	—	13	13	15	—	*14*
13	—	12	—	12	—	14	—	12	12	14	—	*13*
12	—	11	—	11	—	13	—	11	11	13	—	*12*
11	—	10	—	10	—	—	—	10	10	12	—	*11*
10	—	9	—	9	—	—	—	9	9	—	9	*10*
9	—	7	8	7	—	—	—	8	8	—	8	*9*
8	—	6	7	6	—	—	—	7	7	—	7	*8*
7	—	5	6	5	6	—	—	6	6	—	5	*7*
6	—	4	4	4	5	5	—	5	5	—	4	*6*
5	—	3	3	3	4	4	—	4	4	—	3	*5*
4	—	2	2	2	3	3	—	3	2	3	2	*4*
3	—	1	1	1	2	2	—	2	1	2	0	*3*
2	—	0	0	0	1	1	—	1	0	1	0	*2*
1	—	0	0	0	0	0	0	0	0	0	0	*1*

Note. The italicized columns on the extreme left and right of the table represent the age-graded scaled scores that would be obtained from the WAIS–R manual in the standard fashion. The figures under each of the subtests represent the transformed age-graded scaled score (rounded to the nearest integer) for all potential values of the untransformed score.

For ease of use, entries have been left blank where the original and transformed scores do not differ.

between the variances of VIQ and PIQ *within* the UK sample. This revealed that the variances of the scales did not differ significantly (F(199) = 1.09, p = .2667). The question of whether there are significant differences among the *subtest* variances has already been addressed by the test for heterogeneity of variance which preceded the repeated measures ANOVA. It can be seen from Table 5 that the value of $-F_{max}$ does not exceed the critical value, thereby indicating that there is insufficient evidence to reject the null hypothesis that the population variances all have the same value.

Reliabilities of the IQ scales and subtests

The subtest reliability coefficients and SEMs obtained in the UK sample are presented in columns 1 and 3, respectively, of Table 7. For comparison purposes, the reliability coefficients and SEMs obtained from the US standardization sample are presented in columns 2 and 4. The rightmost column of this table contains the SEMs obtained for the UK subtests following a T-score transformation. These values will be discussed in a later section. It can be seen from Table 7 that the majority of the UK subtests have high reliability and also that, in most cases, the reliabilities are very similar in magnitude to the US values.

As noted, the reliabilities of the IQ scales were derived from the reliabilities of the individual subtests using Mosier's formula (Mosier, 1943). It can be seen from Table 8 that, in the UK data, the IQ scales have extremely high reliability and that the coefficients are essentially indistinguishable from those of the US scales.

T-score transformation of the subtest scores

The significant differences among the subtest *mean* scores poses problems for the applied psychologist who is attempting to interpret *an individual's* subtest profile. However, it is a simple matter to rescale the UK subtests so that, in common with their US equivalents, they have a common mean of 10 and a common SD of 3. Such a rescaling was carried out with the present data using T-score transformations. One way of presenting this information would have been to list the relevant equations for the 11 subtests. However, it was considered that a table that set out the transformed values for each of the 11 subtests would be more convenient to use in practice. Table 9 presents the transformed scores for the VIQ and PIQ subtests, respectively.

Discussion

The present results suggest that, in general, the WAIS—R has robust psychometric properties, and therefore provide some reassurance for those psychologists in the UK who are justifiably uneasy about using a test that has never been standardized here. The values of the estimates of the WAIS—R population parameters in the UK are notable more for their similarities to their counterparts in the US standardization sample than for their differences. This is especially true of the three IQ scales: the FSIQ, the VIQ and the PIQ. The findings relating to these three composite scales will be discussed before the individual subtests are considered.

The results of the Kolmogorov—Smirnov tests indicate that the FSIQ, the VIQ and the PIQ have distributions that correspond closely to a normal curve. This is encouraging for users of the UK version of the WAIS—R since, had the distributions been skewed, or leptokurtic/platykurtic, there would have been problems with the interpretation of scores, whether of groups or of individuals.

The estimates of the UK population means for all three scales are very close to the desired value of 100; although, in all three cases, the sample mean is greater than 100. The discrepancies however are sufficiently small that, for all practical purposes, they can be ignored. It should also be noted that (despite the non-significance of the chi-square value) there is a slight over-representation of Social Classes 1 and 2 in the UK sample (see Table 1). In view of the well-corroborated correlation between IQ and social class (in the present sample the Pearson correlation is .60), marginal inflation of the estimate of the population value is to be expected. Furthermore, the population mean in the contemporary US population may now also be greater than

100, because of the phenomenon of IQ gains (Flynn, 1984, 1987). Flynn has demonstrated that, at least throughout the Western World, measured IQ is continuously rising, so that more recent normative samples consistently outperform their predecessors. The WAIS—R was standardized in the US in 1980, over 10 years ago. Therefore, if the trend identified by Flynn has continued since his review, it can be expected that a representative sample of the US population would obtain a mean score greater than 100.

The examination of the reliabilities of the IQ scales also provided encouraging results. The reliability coefficients obtained were exceptionally high and establish the WAIS—R as one of the most reliable psychological tests in use in the UK. Moreover, the coefficients were very similar (indeed, in the cases of the FSIQ and the PIQ, virtually identical) to those obtained from the US standardization sample.

As previously noted, in practical applications of the WAIS-R, analysis is rarely limited to a comparison of the scores of an individual or group with the relative population means. The *profile* of scores is also routinely examined. For this reason, it was important to establish whether or not the VIQ and PIQ scales are comparable in the UK. The present results suggest that the two scales have essentially the same mean and SD and that the distribution of VIQ—PIQ discrepancies is also normal. Although it would still have been possible to examine VIQ—PIQ discrepancies in UK subjects even if the scales did not have these characteristics, the fact that they do greatly simplifies the interpretation of an individual's (or group's performance).

In contrast with the close correspondence between the foregoing UK statistics and the US parameters, significant differences were obtained when the variances of the scales were compared. The present results suggest that the WAIS—R has reduced variance in the UK.

The applied psychologist or researcher should be aware of this finding when comparing the scores of an individual (or group) with the population mean. For example, in the aforementioned BPS (1991) report on mental impairment, it was suggested that an FSIQ that was more than two SDs below the mean could be taken (for legal purposes) as indicating severe mental impairment. In terms of the US standardization, an IQ of 70 would indeed be the cut-off point. The present results, however, suggest that a cut-off point of 74 would be more appropriate for the UK (assuming that the population mean is 100).

The reduced variance also has implications for the construction of confidence intervals for an individual's score. Thus, although the reliability coefficients are essentially equivalent for the US and UK samples, the UK SEMs for the IQ scales are smaller because of the smaller SDs. SEMs for the UK are presented in Table 8.

The statistics of the scores on the 11 WAIS—R subtests give more reason for concern than those of the scores on the three IQ scales. Although, as noted, the reliabilities of the individual subtests were adequate and consistent with the corresponding US values in the majority of cases, there were exceptions to this. The reliability results for three of the subtests justify individual comment. Firstly, it can be seen that the reliability of Arithmetic is substantially lower in the UK sample than in the US standardization sample. It can also be seen that, although the UK reliability of Object Assembly is very similar to the US value, in both samples the reliability of this subtest is relatively poor. Finally, the reliability coefficient for Digit Span is substantially higher in the UK sample. Unfortunately, this result is difficult to interpret, since the method of estimating reliability differed between the two samples. The observed difference may be the result of the difference in method, rather than reflecting a true difference between the samples. From hindsight, it would have been better to use the test—retest method with the UK sample, in order to ensure comparability.

A particularly important finding of the present study is the large (and highly significant) differences among the subtest means. This would suggest that there are in-built discrepancies among the WAIS—R subtests when used in the UK. In the UK, therefore, the analysis of an individual's profile of strengths and weaknesses at the subtest level without a knowledge of these in-built discrepancies is liable to lead to erroneous conclusions. For example, the WAIS—R manual provides a table (13) for use in the individual case, which gives the size of the minimum significant discrepancy between any two subtests. Examination of this table reveals that differences of around two score points are sufficient for significance according to Wechsler's criterion; and yet, in the UK, differences of this magnitude will, in some subtest comparisons, be the norm, rather than the exception. (There are other problems with the aforementioned WAIS—R manual table 13 and its rationale which go beyond the issues connected with its use in the UK: see Crawford [1992] for a discussion.) To combat the problems that the unequal means and reduced variances of the subtests pose for UK users, the simple solution adopted here is to provide a table (Table 9) that can be used to convert US-derived subtest scores to scores with a *common* mean and SD of 10 and 3, respectively, in the UK. Following transformation of the subtest scores, their new SEMs were computed. (The rescaling has the effect of increasing the SEM in every case, since the untransformed SDs were all less than 3.) These SEMs are presented in the rightmost column of Table 7. Where the intention is to examine subtest scatter, it is suggested that clinicians base their interpretation on both the US age-graded scaled scores *and* the T-score converted UK age-graded scores; caution should be exercised in cases where the two sets of scores do not provide consistent profiles.

Before leaving the subject of the subtest means, some comment on the Digit Span subtest is in order. This subtest had the highest mean score (11.36) and the largest deviation, both from the US mean (10) and the grand mean of the UK subtests (10.46). Given the cultural and educational differences between the two countries, it would have been surprising if no significant differences among the subtest means had been found. Digit Span, however, might be thought the most 'culture fair' of the subtests, and so the least likely to show a difference. That plausible assumption, however, may be false. Since modern theories of working memory (e.g. Baddeley, 1986; Baddeley & Hitch, 1974) assign a crucial role to a process akin to repeated articulation (the ***articulatory*** *loop)* in the short term retention of material, memory span for sequences of pronounceable items could be expected to be positively correlated with articulation rate, whether the items are words or digits. There is evidence that this is indeed the case. For example, Ellis & Hennelly (1980) found that Welsh subjects, who were bilingual in English and Welsh, took longer to read sequences of random numbers in Welsh (385 ms per digit on average) than in English (321 ms). On the Welsh and English language versions of the Digit Span subtest of the

Wechsler Intelligence Scale for Children (WISC), the mean score was 6.6 for the English version and 5.8 for the Welsh. Differences in articulation time may also account for the differences in digit span among other linguistic groups (Naveh-Benjamin & Ayres, 1986). In descending order of magnitude of their mean Digit Spans, the four linguistic groups studied by those authors were: English, Hebrew, Spanish and Arabic. However, in terms of digit articulation time, the order was reversed: the longer the articulation time, the lower the Digit Span, so that the English-speaking group achieved the highest Digit Span, the Arabic-speaking the lowest. The articulation times for Chinese digits are even shorter than for English digits; and Chinese subjects have achieved the longest Digit Spans of any of the other groups studied (Hoosain & Salili, 1988). Substantial negative correlations between Digit Span and articulation time have also been found within particular linguistic groups: e.g. Hoosain (1982) found a correlation of —.70 between Digit Span and articulation time with Cantonese subjects; and Naveh-Benjamin & Ayres (1986) report an average within-group correlation of about —.5 over the linguistic groups they studied. Ellis & Hennelly (1980) suggest that different dialects within a language could also give rise to differences in Digit Span. The aforenoted differences in Digit Span among various linguistic groups, therefore, may be paralleled among different regions of the English-speaking world. Since the American English in some states is characterised by a drawl, which reduces the articulation rate, the Digit Span there may be shorter.

It is therefore suggested that the discrepancy between the UK and US means for Digit Span can be attributed to differences in the average articulation rate for digits. This suggestion is tentative; however, it is supported by the fact that the Arithmetic subtest yielded the second-largest difference because mental arithmetic performance has also been found to correlate with articulation rate (e.g. Hitch, 1978).

In conclusion, the present results suggest that, in many respects, the WAIS—R has acceptable psychometric properties in the UK, despite never having been standardized here. Of particular note are the very high reliabilities of the IQ scales and the extent to which these match those of the US standardization sample. Some important differences, however, also emerged, and for this reason it would be valuable (though arduous) to attempt cross-validation of the present results with other sizeable samples drawn from elsewhere in the UK. Clinicians using the WAIS—R in the UK should, we believe, take careful note of the present findings, since they provide the only currently available estimates of the UK population parameters for this test. It should be stressed, however, that studies such as the present one can never be as satisfactory as a formal standardization of the scale in a large, stratified sample. It is to be hoped that a standardization will be carried out in the UK in the not-too-distant future.

Acknowledgement

This study was partly supported by the Mental Health Foundation and Julia Farr Foundation.

References

Anastasi, A. (1988). *Psychological Testing,* 6th ed. New York: Macmillan.
Baddeley, A. D. (1986). *Working Memory.* Oxford: Oxford University Press.
Baddeley, A. D. & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The Psychology of Learning and Motivation,* vol. III. New York: Academic Press. British Psychological Society (1991). Mental impairment and severe mental impairment. *The Psychologist,*
4, 373-376. Crawford, J. R. (1992). Current and premorbid intelligence measures in neuropsychological assessment.
In J. R. Crawford, D. M. Parker & W. McKinlay (Eds), *A Handbook of Neuropsychological Assessment,*
pp. 21-49. Hove: Erlbaum. Ellis, N. C. & Hennelly, R. A. (1980). A bilingual word-length effect: Implications for intelligence
testing and the relative ease of mental calculation in Welsh and English. *British journal of Psychology,*
71, 43-52. Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological bulletin,* **95,**
29-51. Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological bulletin,*
101, 171-191.
Guilford, J. P. (1954). *Psychometric Methods,* 2nd ed. New York: McGraw-Hill. Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology,*
10, 302-323. Hoosain, R. & Salili, F. (1988). Language differences, working memory and mathematical ability. In
M. M. Gruneberg, P. E. Morris & R. N. Sykes (Eds), *Practical Aspects of Memory: Current Research and Issues,* vol. II. New York: John Wiley. Jaccard, J., Becker, M. A. & Wood, G. (1984). Pairwise multiple comparison procedures: A review.
Psychological bulletin, **96,** 589-596. Kaufman, A. S. (1983). Test review of the Wechsler Adult Intelligence Scale-Revised, *journal of*
Psychoeducational Assessment, 1, 309—319. Lea, M. (1986). *A British Supplement to the Manual of the Wechsler Adult Intelligence Scale-revised.* San
Antonio: Psychological Corporation. Matarazzo, J. D. & Herman, D. O. (1984). Base rate data for the WAIS-R: Test-retest reliability and
VIQ—PIQ differences, *journal of Clinical Neuropsychology,* 6, 351—366. Matarazzo, J. D. & Herman, D. O. (1985). Clinical uses of the WAIS-R: Base rates of differences
between VIQ and PIQ in the WAIS-R standardization sample. In B. Wolman (Ed.), *Handbook of Intelligence: Theories, Measurement and Applications,* pp. 899-932. New York: Wiley. Mosier, J. I. (1943). On the reliability of a weighted composite. *Psychometrika,* 8, 161-168. Myers, J. L. (1979). *Fundamentals of "Experimental Design,* 3rd ed. Boston: Allyn & Bacon. Naveh-Benjamin, M. & Ayres, T. J. (1986). Digit span, reading rate and linguistic relativity. *Quarterly*
journal of Experimental Psychology, 38, 739-751.
Nunnally, J. C. (1978). *Psychometric Theory,* 2nd ed. New York: McGraw-Hill. Office of Population, Censuses and Surveys (1980). *Classification of Occupation.* London: HMSO. Saville, P. (1971). *A "British Supplement to the Manual of the Wechsler Adult Intelligence Scale.* Windsor:
NFER-Nelson.
Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill. Tukey, J. W. (1953). *The Problem of Multiple Comparisons.* Unpublished manuscript, Princeton
University. Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological
Corporation. Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-revised.* New York: Psychological
Corporation.

deceived 27 April 1993; revised version received /2 October 1993