Article

Preliminary Validation of a Free-to-Use, Brief Assessment of Adult Intelligence for Research Purposes: The Matrix Matching Test

Graham Pluck [iD]
Quito Brain and Behavior Laboratory, Universidad San Francisco de Quito,
Quito, Ecuador

Abstract
The concept of intelligence as a measurable trait of intellectual function continues to be an important issue in psychology. Traditionally, a core field of differential psychology and widely employed in applied settings, it is also important in various research fields. Here, I describe development of a new assessment of general intelligence of adults that has no language component and can be administered in about 10 minutes. A total sample of 176 adult participants, from various settings, was assessed with a set of matrix tasks that involved either visuospatial (fluid) or semantic (crystallized) reasoning. The internal consistency was acceptable ($\alpha = .748$), and there was good four-week test–retest reliability ($r = .931$). Concurrent validity was demonstrated by a high correlation between the new test and the (seven-subtest version) Wechsler Adult Intelligence Scale-IV (WAIS-IV) scores ($r = .889$). A principal component analysis also suggested that the new test measures the same latent construct as the WAIS-IV—thought to be general intelligence. Predictive validity was shown in a subsample of 60 undergraduates by a medium-sized correlation between test scores and grade point average data ($r = .396$). These preliminary results suggest that the Matrix Matching Test may be a useful research tool.

Corresponding Author:
Graham Pluck, College of Social Sciences and Humanities, Universidad San Francisco de Quito, Diego de Robles y Vía Interoceánica, Cumbayá, Ecuador.
Email: graham@gpluck.co.uk

Introduction

The concept of intelligence as a measurable aspect of the human mind has been with psychology since at least the developments of the first intelligence tests in France in the early 20th century (i.e., Binet, 1903). The very concept of intelligence as a quantitative measure is often controversial among the general public, and even among psychologists. Nevertheless, a statement published in the *Wall Street Journal* in 1994 (and republished in the journal Intelligence in 1997), signed by 57 experts on intelligence, asserted that intelligence can be defined and can be measured and is perhaps in practice the most accurately measured psychological trait (Gottfredson, 1997). That review also provides a useful definition of intelligence: "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (Gottfredson, 1997, p. 13). Of course, being definable does not make something into a real trait. However, the positive manifold, as it is known, is strong evidence that intelligence is a true trait of the human mind. This is the observation that scores on most cognitive tests, perhaps all, are positively correlated with each other, suggesting that they are all influenced by a basic general ability of the human brain for cognitive processing, known as general intelligence (Deary, Penke, & Johnson, 2010).

In recent years, there has been a widening interest in the concept of intelligence. Previously focused on psychometrics and differential psychology, there is increasing realization that intelligence as a concept is a core topic in cognitive psychology (see, e.g., Duncan, Chylinski, Mitchell, & Bhandari, 2017). This is particularly so in cognitive neuroscience which has made significant progress in recent years on the distinctions between executive functions, working memory and fluid or general intelligence, and their potential biological bases (Barbey et al., 2012; Duncan, 2010; Hampshire, Highfield, Parkin, & Owen, 2012). Artificial intelligence, also a rapidly progressing science, is benefiting from understanding human intelligence (Lovett & Forbus, 2017).

Psychometric intelligence is generally measured with commercially developed and normed intelligence tests such as the Stanford-Binet Intelligence Scales (Roid, 2003) or the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2012). Such tests are often appropriate for clinical and educational assessment but present several drawbacks for use in basic psychological research. One problem is the administration time, which can be up to 2 hours. That may be too long for many research protocols. Such extensive assessment is needed to provide precision, which is important when the information may be used for potentially life-changing decisions such as diagnoses. However, for research studies involving comparisons between groups, or correlations with other variables, such precision is often not necessary and may unduly slow the research process. Furthermore, standardized intelligence tests are normed on large samples,

again the normalized scores are often not necessary for research involving group comparisons or correlations.

The high-level of development of commercial intelligence tests, including the norming, also makes them expensive. This can be a real problem for much small-scale research, particularly in less developed countries, where research is often conducted without any financial support. Even if funds are available, researchers often cannot access the tests because there is no sales infrastructure. Some researchers, and most regrettably clinicians too, resort to using poorly photo-copied versions of commercial tests, which may invalidate the measurements.

Even if full commercial intelligence tests are purchased and used, there are often problems with their applicability outside of the countries that they were developed in. The most obvious issue is with language. Large intelligence bat-teries include assessments of verbal ability. This is appropriate as there is evi-dence that general intelligence, which psychometric intelligence tests such as the WAIS-IV are strong measures of (Canivez & Watkins, 2010), has at least two facets, crystalized and fluid intelligence. This distinction is based on the classic work of Raymond Cattell (1967, 1973) but seems to hold true in recent analyses (Benson, Hulac, & Kranzler, 2010; Hampshire et al., 2012; Schipolowski, Wilhelm, & Schroeders, 2014). In the words of Cattell (1967), crystalized intel-ligence involves "skilled judgment habits (that) have become crystallized . . . as a result of earlier learning," while fluid intelligence involves "adaption to new situations where crystallized skills are of no particular advantage" (pp. 2–3). Fluid intelligence assessment can be relatively "culture fair" (Cattell, 1973) and has significant overlap with concepts of executive function and working memory (Conway, Kane, & Engle, 2003; Engle, Tuholski, Laughlin, & Conway, 1999). On the other hand, crystalized intelligence appears to be dependent on both language ability and general knowledge (Schipolowski et al., 2014). It is therefore often measured with tests such as the WAIS Vocabulary, Similarities, and Information subtests. This presents particular problems for use of comprehensive intelligence tests such as the WAIS outside of the countries that they were developed within, as all of those tests are heavily dependent on spoken language.

An alternative of course is to use assessments that rely only on fluid intelli-gence measurement, particularly as these tests usually involve visuospatial rea-soning and have no overt language component. This is a common strategy and the one embodied in two common intelligence tests, the Raven's Progressive Matrices (Bingham, Burke, & Murray, 1966; Raven, 1936) and the Cattell Culture Fair (Cattell, 1973). Nevertheless, by neglecting to measure crystalized aspects of intelligence, the measurement could be seen as incomplete.

Furthermore, even the subtests that appear to measure fluid components of intelligence often have aspects that are culture dependent. For example, one of the tasks in the WAIS-IV requires participants to spontaneously attend to the accumulation of snow on logs in a rural log cabin scene. In many countries,

people have never experienced snow and so there are issues about the validity. Although this may seem overcritical, the point is not that commercial intelligence tests are at fault, but merely that they have problems when applied beyond their intended uses.

If we examine other areas of research, such as cognitive psychology and experimental neuropsychology, most of these problems generally do not exist. This is because there is an academic culture of developing assessment methods as open-access tools, which often can be supplied by the test authors free of charge, downloaded, or even constructed from the instructions in the journal articles. They can then be adapted to the local needs. As an example, the website of Cambridge University (UK) professor Simon Baron-Cohen (www.autismresearchcentre.com) contains various tests (in various languages) for studying autism and theory of mind that can be downloaded and used in research without cost. Although research tools downloaded from the Internet may not meet the high standards required for clinical or education assessments, they have much potential to promote further research.

Here, I present initial development of a brief test of general intelligence that can be downloaded and used by researchers on an experimental basis without cost. The files can be downloaded from http://www.gpluck.co.uk/Tests.html. The test, called the Matrix Matching Test, is administered as PowerPoint files on a tablet computer, making it highly portable and involving no costs other than for the tablet computer. On face validity, the test accesses both fluid and crystallized intelligence to give an estimate of general intelligence. This is achieved by inclusion of visuospatial matrix reasoning tasks, as well as tasks in which, to complete the matrix, participants must match items based on semantic understanding. The test has been created as much as possible to be culture independent and there is no overt language component. There is no attempt to provide normative data, but I have assessed the reliability and validity, giving a starting point for other researchers to potentially use the test in their own research.

Methods

Participants

Five different samples were recruited and used for different parts of the piloting, reliability, and validity analyses of the Matrix Matching Test. All were Spanish speaking adults.

The first sample was the "Pilot" sample. This sample was used to pilot the initial set of matrices and was composed of 64 undergraduate students from various majors at Universidad San Francisco de Quito (Ecuador). The majority of this sample (47/64, 73.44%) were female and the mean age was 19.85 (SD = 2.27).

The second sample, the "University sample," was recruited primarily from employees of Universidad San Francisco de Quito (Ecuador) and from friends

and family of people who had already participated. The aim was to recruit adults from a range of backgrounds so as to avoid limitations of range in the data, which tends to happen when specific groups are targeted. Fifty participants were recruited in this way. Although five undergraduate and two postgraduate students were included, the majority of the participants were non-students working in a range of occupations. The most common categories were cleaner ($n = 10$) and security guard ($n = 8$), the rest had various occupations in education, administration, technical, and maintenance. Two participants described themselves as unemployed and two as retired. The mean age in years of the sample was 35.24 (SD = 11.37, range = 18–65) and the mean years of formal education was 14.32 (SD = 3.94, range = 6–26). The majority, 27/50 (54.00%), were women. This sample was recruited mainly to examine the correlations between the Matrix Matching Test and a standard scale of intelligence, the WAIS-IV.

The third sample was called the "Test–retest sample." There were 21 participants recruited from the same population as the University sample, i.e., students or staff at Universidad San Francisco de Quito. However, most of these were not assessed with the WAIS-IV as the objective was to examine the test–retest reliability of the Matrix Matching Test. However, eight of the participants in this sample later participated in the University sample described above. The eight who participated in both samples were five people working as security guards and three as cleaners. The 13 who participated only in the Test–retest sample were undergraduate students (n = 5), university professors (n = 2), security guards (n = 2), and various other occupations. The mean age of the whole Test–retest sample of 21 participants was 33.67 (SD = 10.61, range = 18–53) and the mean years of education was 13.52 (SD = 4.24, range 7–26). There were fewer women than men in this sample, 9/21 (42.86%).

The fourth sample collected was of 60 students at a different university in Ecuador (Universidad Nacional de Chimborazo). The student sample was collected to examine the predictive validity of the Matrix Matching Test. This group of participants was called the "Student sample." All 60 students were studying for undergraduate degrees in engineering and had completed at least three semesters of study. The mean age of the students was 23.52 (SD = 3.05, range 20–37) and only 14/60 (23.33%) were women. This sample was recruited as part of a different research study on neuropsychological predictors of academic achievement.

The fifth and final sample was composed of new-car sales personnel and is henceforth referred to as the "Sales sample." Fifty-three individuals were recruited as part of a separate study on prediction of work place achievement with neuropsychological measures. Their data are included here to improve the accuracy of the internal consistency estimates of the Matrix Matching Test and other correlations. The mean age of this sample was 37.42 years (SD = 7.09, range = 21–54) and the mean years of education was 15.13 (SD = 1.90, range = 12–19); 20 (37.74%) were women.

Development of the matrix matching test

An initial pilot version of the Matrix Matching Test was developed which comprised two subtests: One of 18 visuospatial reasoning tasks and the other of 23 semantic reasoning tasks. The visuospatial tasks all comprised a matrix of similar visual stimuli which followed a pattern. The different elements in the matrix varied systematically with change over one or several dimensions (e.g., color, orientation, and size). This manipulation of the number of dimensions of change influences the difficulty. In each of the matrices, one element was missing. Six possible elements were given in a panel below and the task is to pick the one that completes the matrix. When the matrix elements vary by only one dimension, it is relatively easy to spot the correct piece that is missing, but this is more difficult when several dimensions vary systematically. The visuospatial reasoning tasks are in color and contain no verbal material or recognizable objects (other than circles, lines, squares, etc.). One point is awarded for each correctly solved. These visuospatial matrix tests are similar in concept to those used in several existing "culture free" general intelligence tests such as Raven's Progressive Matrices (Bingham et al., 1966; Raven, 1936).

The pilot tasks in the semantic reasoning set use a similar format to the visuospatial set; however, actual images, usually photographs, are used. In each, the task is to detect the concept that links all the items in the matrix and pick the one from an array below that completes the set. For example, in one task, there are several flightless birds in the matrix (images of a kiwi, an ostrich, and an emu). The task is to pick the one from the example array given below which completes the set so that they all show the same concept; in this case, the correct response is an image of a penguin, as it is the only flightless bird in the array (the distractors are images of a crow, a polar bear, a cat, and a brown bear). Note that this could also be interpreted as being of southern hemisphere creatures, but the correct response is the same. Some of the later tasks have two elements missing from the array and the task is to pick the two from the possible elements which maintain the semantic set. This allows more specific focus on concepts and minimizes participants choosing elements that have some link to the elements in the matrix other than intended by the researcher. As an example, in one such task, the target stimuli are images of ice cubes, clouds, and a glass or water. The participant then must choose both the image of rain and the image of steam to complete the set of "forms of water" (the distractors are smoke coming from a chimney, an empty drinking glass, and a gazelle). One point is awarded for set completed correctly (maximum one point per trial).

All the images in the semantic reasoning set were taken from Google Images and were marked as available for non-commercial reuse. Where people occur in the images, these represent a variety of different races, including people of White, Black, and Asian appearances. In addition, images and concepts were deliberately selected to minimize effects of variation of geographical or

socioeconomic exposure. Images of things which would be likely familiar to anybody living in at least somewhat industrialized countries were used. For example, images of various forms of transport, including cars, motorbikes, and airplanes were included as these are common in most cultures (or at least people have seen images of them). Similarly, food stuffs such as bananas or rice were used as these are commonly found around the world at all socioeconomic levels. Concepts that more obviously vary geographically, such as aspects of religion or the weather were avoided. However, with such a test, it is impossible to be completely culture-free. Although this test has the same format as the visuospatial reasoning tasks, conceptually it has similarities to the Pyramids and Palm Trees Test which is used to assess access to semantic information in neuropsychological patients (Howard & Patterson, 1992). In that test, triplets of concepts are shown to patients, such as an image of a pyramid, with two examples that may match it in some way, such as a palm tree and fir tree. The patient is asked to pick which of the trees matches the pyramid image best.

Both the visuospatial and semantic tasks are also free of language that links them to a particular culture. There are no Arabic numerals to identify choices. Instead numbers are represented by quantities of dots. This is used to reduce cultural bias. All of the tasks were produced in Microsoft PowerPoint so that they can be displayed on tablet computers. Recording of responses is by hand.

These 41 pilot matrices (18 visuospatial and 23 semantic) were initially tested on the Pilot sample (described above in the Participants section). The piloting was conducted as part of a separate study on neuropsychological predictors of academic achievement, where further details of the sample can be viewed (Pluck, Ruales-Chieruzzi, Paucar-Guerra, Andrade-Guimaraes, & Trueba, 2016). This pilot data were used to remove items which were ambiguous or were uncorrelated with standardized intelligence test scores and to alter some of the items. From this first stage, a second still-in-development version containing 16 visuospatial and 16 semantic items was produced. This in fact comprised 14 items of each type plus two additional of each type included on an experimental basis. Data collection began with the University sample. However, preliminary analyses revealed that the additional two items of each type did indeed seem to perform poorly and so were dropped from the studies involving the other samples. The majority of the participants reported below therefore only completed a 28-item version comprising 14 visuospatial and 14 semantic trials. The data presented in the initial part of the results are based on those 28 trials that together comprise the early version of the Matrix Matching Test and were completed by all participants.

Standardized intelligence assessment

As part of the validation of the Matrix Matching Test, we also assessed the participants in the University sample with a "gold-standard" intelligence test.

For this, we used the Spanish-language version of the WAIS-IV (Wechsler, 2012). This is the most widely used intelligence assessment in Spanish-speaking countries. It was normed in Spain and published in 2012. The full WAIS-IV requires the application of 10 subtests; however, for brevity, we employed a standardized shorter seven-subtest version, comprising Block Design, Similarities, Digit Span, Arithmetic, Information, Coding, and Picture Completion subtests. This abbreviated version has been shown to measure intelligence practically as well as the full version; in fact, the correlation between the seven-subtest and full administrations is $r = .99$ (Meyers, Zellinger, Kockler, Wagner, & Miller, 2013). The validity of this version in Ecuadorian samples has previously been established (Pluck et al., 2016). Although the seven-subtest version does not contain the Vocabulary subtest, often considered an excellent measure of verbal intelligence, it does contain two other subtests thought to measure verbal ability—Similarities and Information. The seven-subtest WAIS-IV takes around 40 minutes to administer.

Procedure

Prior to the start of data collection in all samples, participants signed a consent form, in accordance with the ethics committee approvals for the different studies. In all samples, basic demographic data were collected such as age and educational experience and then the cognitive tests were administered. The Matrix Matching Test was administered as PowerPoint files on 10-inch Galaxy tablet computers. Responses were recorded by hand. Participants progressed through the individual trials of first the Visuospatial and then the Semantic subtests. There was neither time limit on individual trials nor any termination rule: All participants attempted all trials. Feedback on performance was not given, except when participants made errors in the first three trials of either the Visuospatial or Semantic subtests, which in practice rarely occurred. The individual trials were sequenced from easy to difficult, so if any errors were made on the first three trails, it may have indicated a misunderstanding of the task and so an explanation of the correct response and the reason that it is correct were given. Administration of the Matrix Matching Test took around 10 minutes.

The participants in the University sample and the Test–retest sample were all recruited and assessed at Universidad San Francisco de Quito in Ecuador; this is a private university located near the city of Quito. Recruitment was by posters and by word of mouth, individuals who had already participated sometimes told their friends about the study, who then later participated. They were assessed by research assistants under the supervision of the author. All assessments were performed in a private interview room. For the University sample, the full interview was conducted in a single session. All participants were assessed with the

32-item version of the Matrix Matching Test. They were also assessed with the WAIS-IV. A few other tests were administered in this session, which are not reported here. The entire session took around 90 minutes to complete. The participants were then debriefed and given a payment of US$20. Relatively large financial incentives were used to attract participants from a range of socio-economic backgrounds.

For the Test–retest data, participants were all assessed by the same one researcher. They were assessed twice on the same tests with a 4-week delay between assessments. The test and retest assessments occurred at the same time of day where possible, and always in the same location. This assessment included the Matrix Matching Test (28-item version) and a small number of other tests not reported here. Each test session took around 20 minutes. None of the participants was assessed with the WAIS-IV at either test or retest. Although 21 participants were recruited, one did not return for the retest assessment; therefore, the test–retest statistics are on a sample of 20 participants. After the completion of the second assessment, all participants were debriefed and given a payment of US$20.

Eight participants from the Test–retest sample were then recruited again into the University sample. This effectively meant that they returned for a third visit in which they were assessed with the WAIS-IV, which took around 40 minutes. In this group of eight, the data that were used for analysis was the Matrix Matching Scores from the first visit on test–retest and their later WAIS-IV scores.

For the sales personnel, all participants were recruited at their place of work, which were several different Chevrolet dealerships in the city of Quito, Ecuador. They were all assessed by qualified psychologists. Interviews were conducted in offices at the workplaces. The Matrix Matching Test (28-item version) was administered as part of a larger battery of cognitive tests not reported here. The entire assessment took around 90 minutes. Participants were thanked and debriefed but were not paid a financial incentive.

The Student sample was recruited in Universidad Nacional de Chimborazo, a state-run university in the city of Riobamba, Ecuador. All were undergraduate students of engineering who had already completed at least three semesters of study. They were assessed by professors of the Department of Psychology and Education. All assessments were conducted in quiet private rooms at the University. Again, the Matrix Matching Test (28-item version) was administered as part of a larger battery not reported here. The entire assessment took around 90 minutes. Participants were then debriefed and received course credit for their assistance with research. Later, data on each student's academic performance were taken from the University computer systems. The Grade Point Average (GPA) for each student for the whole semester following the semester in which they participated was recorded.

Results and discussion

Internal consistency and score distributions

For this analysis, data from all four samples (excluding the Pilot sample) were included. These were analyzed in a combined sample containing all of the participants (N = 176). The common 28 trials from the Matrix Matching Test were analyzed together as the intent is to use them in a single scale of intelligence. To estimate the internal consistency of this proposed intelligence scale, we examined the Cronbach's alpha (α) values. The initial α with all 28 items was 0.732. However, we also examined the α values when individual items were removed. This showed that removal of two visuospatial and two semantic items improved the internal consistency, up to 0.748. However, no further increases in the α value could be made by item removal.

The two visuospatial items that were removed were both late in the task and rather complicated; nevertheless, they were answered correctly by around half of the participants (44.9% and 50.6%). However, they both had quite low item-total correlations (0.170 and 0.147), meaning that they may have been sensitive to a factor other than that of the remaining items in the set. In the first, this might have been caused by a mental rotation element among the response options, a feature not present in the other tasks. In the other, it may have been that two options appeared to be correct but for one small difference, and this may have lured people into making incorrect responses despite good performance on the previous tasks. Of the semantic matching tasks, the first one removed was answered correctly by 49.4% of the participants, but had a very low item-total correlation (0.058). This item had the concept of "living things" and the supposed correct response was to choose an image of a cow to match an array comprising images of a woman, a flower, and a fish (the distractors were all inanimate objects). It is unclear why this had such a low item-total correlation. The final item removed was a semantic matching task in which the concept was gravity. This also had a very low item-total correlation (0.041) which may be related to the difficulty of the task in which only 14/176 (7.95%) of the participants were able to correctly identify the correct responses.

The details of this final version of 24 items are shown in Table 1. We then examined the α values of this 24-item set in the four different samples individually. These values can be seen in Table 2. In the individual samples, the α values were all qualitatively "acceptable" to "good" based on standard interpretations, with the exception of the student sample in which the value would be considered "unacceptable." This is probably due to the lower variance in Matrix Matching Test scores in that sample, as can be seen in comparison of the standard deviations in the different samples. Cronbach's α is composed partly of correlation analyses and is affected by the limitation of range within data sets. Also included in Table 1 are the Cronbach's α values for the subgroups when the full set

Table 1. Response accuracy and internal consistency details of the final 24 items used in the Matrix Matching Test.

Item	% correct	Corrected item: Total correlation	α if deleted
Vs1	99	.100	.748
Vs2	98	.140	.747
Vs3	86	.250	.742
Vs4	86	.341	.736
Vs5	86	.289	.740
Vs6	79	.357	.735
Vs7	62	.268	.742
Vs8	55	.341	.736
Vs9	45	.211	.746
Vs10	43	.415	.729
Vs11	40	.241	.744
Vs12	26	.192	.746
S1	98	.290	.743
S2	97	.183	.746
S3	96	.284	.742
S4	95	.236	.744
S5	78	.346	.735
S6	68	.384	.732
S7	61	.332	.736
S8	56	.480	.724
S9	39	.244	.743
S10	36	.461	.726
S11	31	.248	.743
S12	25	.275	.740

Note: Vs: visuospatial; S: Semantic. The items in each subtest have been reordered into level of difficulty based on the percentage of the sample that responded correctly.

of 28 items is included. As would be expected, these tend to be a little lower than in the restricted 24-item set, although they are all in the same general qualitative categories of "acceptable" to "good" (and again "unacceptable" for the Student sample).

The total scores on the Visuospatial and Semantic subtests were significantly and positively correlated with each other in all of the samples, and with the exception of the Student sample, these correlations would be described in qualitative terms as "large" (Cohen, 1992). This further suggests that combination

Table 2. Psychometric properties, inter-correlations, and correlations with demographic variables of the 24-item Matrix Matching Test scores in the different samples.

	Correlated with	Sales	University	Test–retest	Student	Combined
Sample size		53	42	21	60	176
Cronbach's α (24 items)[a]		.749	.786	.855	.567	.748
Cronbach's α (28 items)[a]		.724	.789	.839	.564	.732
Mean score (+SD)		16.51 (3.54)	16.93 (3.84)	14.76 (4.74)	14.93 (2.97)	15.86 (3.68)
Score range		8–24	9–23	4–22	8–23	4–24
Visuospatial scores	Semantic scores	.548***	.585***	.739***	.305*	.539***
Age	Visuospatial	−.193	−.412**	−.495*	−.185	−.148*
	Semantic	.005	−.130	−.317	−.219	.006
	Total score	−.093	−.294	−.440*	−.248	−.078
Education	Visuospatial	.186	.372*	.567**	N/A	.390***
	Semantic	.189	.641***	.724***	N/A	.524***
	Total score	.213	.579***	.688**	N/A	.513***

[a]The Cronbach's α values for both the final set of 24 items and the earlier development set of 28 items are shown.
*$p < .05$, **$p < .01$, ***$p < .001$.

of the 12 visuospatial and 12 semantic items into a single score is appropriate. The following analyses are based on this 24-item version of the Matrix Matching Test. Analysis of the total scores in the combined sample suggests that the distribution does not deviate from normal (Shapiro–Wilk (176) = .986, $p = .085$). The distribution of Matrix Matching Test scores for the full sample (N = 176) is shown in Figure 1. Visual analysis of the distribution suggests possible outliers; however, this was checked with a standard procedure (Leys, Ley, Klein, Bernard, & Licata, 2013) and all scores fall within an acceptable range (three median absolute deviations from the median). The means and ranges of scores in the different samples are also shown in Table 2. The highest scoring group was the University sample which scored a group mean of 16.93 items correct, this is actually two points higher than the Student sample. The reasons for this are unclear but it may be that the inclusion of some highly educated professors and two postgraduate students in the University sample raised the score somewhat, compared with the wholly undergraduate Student sample. In general, the test seemed to be able to divide individuals based on performance, i.e., provide a large spread of scores between the best and worst performers on the test.
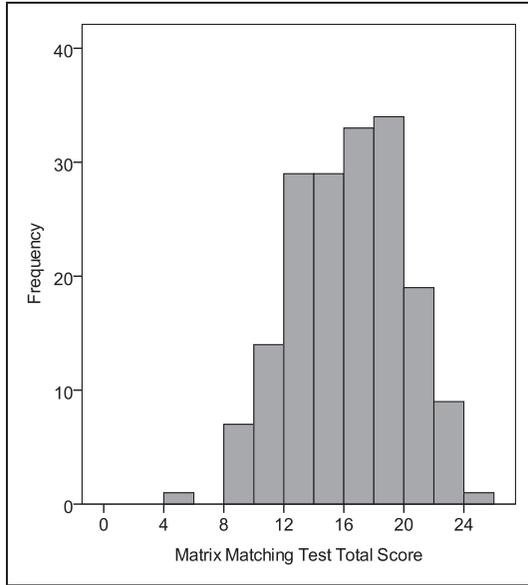
Figure 1. The distribution of scores on the final 24-item Matrix Matching Test for the combined sample of participants (N = 176).

The lowest score was 4 correct and the highest score was 24 correct. Twenty-four is the maximum possible score; however, only one participant from the combined sample of 176 participants scored perfectly and so there is no practical ceiling effect in the data. The distribution of scores suggests that individuals in general adult populations will tend to achieve total scores over an effective range of 21 points (4 to 24 correct). Also suggesting no problem with ceiling or floor effects is that the maximum possible score is 2.21 SDs above the observed mean and the minimum possible score is 4.31 SDs below the observed mean. The greater potential range below what could be expected from a general population mean could be useful if the test were used with clinical groups who would generally show, if anything, lower scores.

Associations with demographic variables

There were no significant differences between male and female participants for the total score on the Matrix Matching Test ($F(1,174) = 0.619$, $p = .433$, $\eta^2 = .004$) nor for either the Visuospatial ($F(1,174) = 0.303$, $p = .583$, $\eta^2 = .002$) or Semantic ($F(1,174) = 0.674$, $p = .413$, $\eta^2 = .004$) subtest components. Although the majority of the Student sample were male, as in common on engineering courses, and engineers might be expected to perform well, they

actually had a slightly lower mean score than the University sample composed mainly of employees.

I also examined how age or years of education correlate with performance on the Matrix Matching Test. Age had a small but significant negative correlation with Visuospatial subtest scores in the combined sample, indicating possible decline in performance with increasing age. However, this varied depending on the group studied, showing larger and significant effects in the University sample and the Test–retest sample, but smaller and not significant effects in the Student and Sales-personnel samples. Nevertheless, in all comparisons, the correlation r values are negative indicating poorer performance in older participants. In contrast, the Semantic subtest scores appeared to show less association with age. Although the trend was in the same direction as for the visuospatial tasks, the correlation r values are somewhat lower, and none reach statistical significance. Overall, the correlations with age suggest generally worse performance in older adults, particularly involving performance on the Visuospatial subtest. This is consistent with the suggestion that the Visuospatial subtest particularly assess fluid intelligence, while the Semantic subtest assess additionally crystalized intelligence. It is known that while fluid task performance declines even with healthy aging, crystalized performance is stable until at least the eighth decade (Park, 2000).

In contrast, there are larger positive correlations between years of education and performance; these were significant in all samples except the Sales-personnel sample. That coefficient may be limited by the variability in education level, which was lowest within that group. These correlations were not performed on the Student sample as their educational experience is more or less the same for each participant. In general, the results suggest that participants with the longest formal educational experiences had the best scores. In all comparisons, whether significant or not, the correlation r values were positive, and again in all comparisons, they were higher for the Semantic subtest than for the Visuospatial subtest. This is as would be expected if the semantic subtest draws more on crystalized ability than the visuospatial subtest.

Test–retest reliability

To assess the stability of the Matrix Matching Test scores over time, 20 participants performed the test once and then again 4 weeks later. The correlation between the two sessions were high for the total 24 items, $r = .931$, $p < .001$, as well as for the Visuospatial subtest scores, $r = .888$, $p < .001$, and for the Semantic subtest scores, $r = .894$, $p < .001$. These suggest good test–retest reliability. However, it also important to know whether there is any trend for systematic change in performance over time, such as a practice effect. The mean score on the full test at first administration was 14.70 (SD = 4.86) and at retest was 16.10 (SD = 4.79), showing an absolute increase of 1.40 points.

The difference was statically significant ($F(1,19) = 12.250$, $p = .002$, $\eta^2 = .392$). The practice effect appears to be primarily driven by improved performance on the Semantic subtest in which there was an absolute test–retest increase in scores of 0.90 points (from 6.95, SD = 2.44 to 7.85, SD = 2.43), which was a significant change ($F(1,19) = 12.933$, $p = .002$, $\eta^2 = .405$). In contrast, the test to retest change in Visuospatial subtest performance was only .50 points (from 7.75 (SD = 2.73) to 8.25 (SD = 2.65)) and that difference was not significant ($F(1,19) = 3.065$, $p = .096$, $\eta^2 = .139$).

Overall, the test–retest analyses suggest that the Matrix Matching Test is measuring a stable construct and is appropriate for single assessments. However, the large practice effect suggests that it may not be appropriate for research in which repeat assessments are performed.

Concurrent validity

To assess the concurrent validity, I examined the correlations between scores on the Matrix Matching Test and scores on the WAIS-IV. The WAIS-IV was only administered to the University sample. As that sample has a large age range (18–65) and the correlation analyses shown in Table 2 suggest that age is associated with performance differentially for the Visuospatial and Semantic subtests, these were age corrected. This is important because IQ scores on the WAIS-IV are age corrected, so the scores on the Matrix Matching Test should also be corrected to allow a fair assessment of the correlation. This was achieved by regressing age on the Visuospatial subtest scores and saving the residuals. The residuals thus represent age-corrected versions of the individual data points. I then did the same with the Semantic subtest scores. To create an age-corrected total score, I simply summed the age-corrected scores from the Visuospatial and Semantic subtests.

The correlation between WAIS-IV full-scale IQ scores and the age-corrected Visuospatial subtest scores was positive and significant, $r = .769$, $p < .001$. WAIS-IV full-scale IQ was also positive and significantly correlated with the Semantic subtest scores, $r = .866$, $p < .001$, and with the total Matrix Matching Test score, $r = .889$, $p < .001$. Interestingly, the correlation is somewhat higher in the Semantic subtest compared with the Visuospatial subtest. This suggests that the addition of semantic reasoning trials improves the ability of the test to assess intelligence. To assess whether the inclusion of the Visuospatial subtest actually adds anything beyond that measured by the Semantic subtest to the prediction of WAIS IQ scores, I performed a linear regression of the two Matrix Matching Test subtests as independent variables as predictors of the WAIS IQ score. Semantic scores were entered first and then Visuospatial scores were entered. The final model including both independent variables had an R of 0.894 (adjusted $R^2 = 0.791$) and the model was a significant predictor of WAIS IQ scores ($F(2,47) = 93.832$, $p < .001$). Importantly, the increase in the R^2 was from 0.750 to 0.800 and this change was significant ($F(1,47) = 11.705$, $p = .001$)

indicating that the Visuospatial tasks do in fact add to the prediction over above that provided by the Semantic task.

This may indicate an advantage of the Matrix Matching Tasks over other intelligence tests which use only visuospatial reasoning, such as Raven's Progressive Matrices. Indeed, the Semantic subtest scores correlate with WAIS-IV full-scale IQ almost as well as the total Matrix Matching Test score. It could be construed that even the Visuospatial subtest is unnecessary. Nevertheless, there is some improvement in the correlation with inclusion of the visuospatial items, and in addition, the use of the Semantic subtest alone would limit precision, as the potential score range would be from 0 to 12. With the full Matrix Matching Test, the range is from 0 to 24.

Overall, the results suggest that the Matrix Matching Test has good concurrent validity. The overall correlation with WAIS-IV IQ scores of $r = .889$ indicates that there is about 79% shared variance between the two measures ($r^2 = .790$). I also examined how well the Visuospatial and Semantic subtests of the Matrix Matching Test correlate with the seven subtests of the WAIS-IV that were used in this research. For these calculations, I used the age-corrected Matrix Matching Test scores and the age-corrected standard scores from the WAIS-IV manual. The results are shown in Table 3.

All of the correlations were positive, qualitatively "large" effects and significant at $p < .001$. As with the correlations with full-scale IQ scores, the Semantic subtest of the Matrix Matching Test generally has the highest correlations, although for the Visuospatial subtest, these are nonetheless large and consistently positive correlations. Furthermore, the two highest correlations of the Semantic subtest of the Matrix Matching Test with WAIS subtests are for Similarities and Information (both $r > .8$). Both of these tests load most highly

Table 3. Correlation coefficients (r values) between subtests of the WAIS-IV and the Matrix Matching Test.

WAIS-IV subtest	Matrix Matching Test subtest	
	Visuospatial	Semantic
Block design	.591	.684
Similarities	.674	.822
Digit span	.699	.669
Arithmetic	.701	.726
Information	.674	.810
Digit-symbol coding	.677	.766
Incomplete figures	.697	.777

Note: All the correlation coefficients in the above table are significant at $p < .001$.

on a crystalized intelligence factor when WAIS-IV subtests are factor analyzed (Benson et al., 2010).

In contrast, the two highest correlations of the Visuospatial subtest with the WAIS-IV subtests are for Digit Span and Arithmetic. In factor analytic studies, these both load onto a working memory factor (Benson et al., 2010; Canivez & Watkins, 2010). However, the basic underlying cognitive process that distinguishes good from poor performers on classic fluid intelligence tests such as Raven's Progressive Matrices is thought to be the ability to maintain subgoals in working memory (Carpenter, Just, & Shell, 1990), and working memory ability may be one of the essential elements of general fluid intelligence (Kane et al., 2004). This suggests that while the Semantic subtest accesses crystalized-verbal ability, the Visuospatial subtest is more focused on fluid ability. That the Matrix Matching Test accesses both of these features of general intelligence may explain the high correlation observed with full-scale IQ scores.

The Matrix Matching Test therefore may be measuring a similar latent construct to that measured by the WAIS-IV, which is thought to be general intelligence (Canivez & Watkins, 2010). As an exploratory measure to examine this, I performed a principal component analysis with the raw scores of the seven WAIS-IV subtests and the total raw score on the Matrix Matching Test. Although the sample size is low for such analyses, where there is a clear, strong factor structure, sample sizes as low as 50 can be appropriate (Barrett & Kline, 1981). The principal component analysis identified only one component with an eigenvalue greater than 1.0, the actual value of this factor was 5.997, which accounted for 74.99% of the variance. The highest loading test on this component was the WAIS-IV Similarities (0.932), followed by the Matrix Matching Test (0.921), Incomplete Figures (0.876), Arithmetic (0.859), Digit Span (0.850), Information (0.847), Digit-Symbol Coding (0.841), and finally Block Design (0.791). This therefore suggests that the Matrix Matching Test is equally efficient as the subtests of the WAIS-IV in measuring the latent construct that they are measuring, assumed to be general intelligence.

Predictive validity

To assess the ability of the Matrix Matching Test to predict "real-life intelligence," I examined its ability to predict academic performance. GPA data were available on all of the participants in the Student sample. GPA in the university where the research was conducted is scored from 0 to 10, with higher scores indicating better academic performance. There was a significant, medium-sized positive correlation between the overall Matrix Matching Test scores and GPA, $r = .396$, $p < .001$. There were also significant positive correlations of GPA with both the Visuospatial subtest, $r = .310$, $p = .016$, and the Semantic subtest, $r = .333$, $p = .009$. These therefore suggest that the Matrix Matching Test has predictive validity in terms of predicting academic achievement. The correlations

would be considered as medium sized by conventional interpretations; the main correlation of GPA with overall task performance has an r^2 of .157, indicating about 16% of shared variance between the measures. Although this may not seem much, in general, intelligence testing is a fairly poor predictor of performance in higher education and the correlations observed here are in fact higher than is usually observed (Richardson, Abraham, & Bond, 2012).

Finally, as the predictive validity was established on the reduced set of 24 items (four were removed based on the Cronbach's alpha scores), it may be of interest to examine whether this amendment is driving the correlations. This does not seem to be the case, the correlation reported above with the final Matrix Matching Test of 24 items ($r = .396$) was not much higher than the correlation calculated on the full set of 28 items ($r = .382$, $p = .003$). The small difference would be expected as the final 24-item version has better internal consistency. I also explored if the post hoc removal of items might drive the correlation with the WAIS that was used to establish concurrent validity. Again this does not seem to be the case, the original correlation based on the final 24-item Matric Matching Test ($r = .889$) is not much higher than is achieved when the full 28 items are used to calculate the total score ($r = .880$, $p < .001$) and again explicable by the improved internal consistency of the reduced set Matrix Matching Test. In the same idea, four items were removed from the test used with the University sample (who did a 32-item version). The other samples completed the 28-item version. This post hoc change does not appear to have driven the concurrent validity correlation, as in fact the r value would have been slightly higher if those four items had been kept ($r = .861$, $p < .001$).

General discussion

The Matrix Matching Test presented here has adequate internal consistency, test–retest reliability, concurrent validity, and predictive validity. It has the potential to be used as a brief measure of general intelligence for research purposes. In addition to its brevity, the test has other positive features. First, it is free to use. There is no license, and in fact, it can simply be downloaded and used. Second, because it is administered as PowerPoint files on tablet computers, it is highly portable, which can be useful for field research such as in schools or clinics. In its current form, responses are recorded by hand; however, as the basic stimuli files are provide free to use, they could potentially be adapted by users so that responses are recorded automatically.

Although I have presented the general psychometric properties of the Matrix Matching Test, as with any other psychological test, its validation only really holds for the culture it was validated in. In this case, this was in a South American country, Ecuador. However, the test has been deliberately designed to minimize aspects that might invalidate its use in other countries. It contains no overt language component, and images have been chosen such that intelligent

people in other countries should not be disadvantaged by their familiarity with the concepts.

Nevertheless, researchers in other countries may need to reestablish some of the psychometric properties in their own samples. This could include estimation of the internal consistency, if it used with samples containing sufficient variation. When any scale is used with a low-variability sample, it tends to produce data with a restricted range and hence estimates of internal consistency tend to be low and uninformative. In general, the Matrix Matching Test can be used as raw data, based on the total score from the Visuospatial and Semantic subtests. This will usually be appropriate when comparing scores from different groups who are matched on demographic factors such as sex and age. If the raw data is used in correlations, and if there is a wide age range within the sample, then this could affect the results. In such cases, age-corrected scores could be considered. A statistical method for this is described in the Results section. However, this potential need for age correction is true of most cognitive tests. In fact, the Matrix Matching Test may be more robust than other tests of intelligence commonly used in research, i.e., those that rely entirely on assessment of visuospatial reasoning. This is because, as observed in the current study, and by others (e.g., Park, 2000), tests with more crystalized content show less decrement with normal healthy aging than fluid reasoning tasks.

This is one of the other advantages of the Matrix Matching Test, the inclusion of items that likely assess crystalized intelligence, despite being language-free in its administration. In addition to the face validity of the Semantic subtest as a measure of crystalized ability (one has to understand semantic concepts to respond correctly), that it can assess crystalized ability is implied by the qualitatively large correlations with other factors. Specifically, these are the correlations with education level and scores on measures of language and declarative knowledge (WAIS Similarities and Information subtests). Crystalized ability, compared with fluid ability, is theorized to be more dependent on cultural factors, particularly education (Cattell, 1967), while language ability and factual knowledge are the two central components of crystalized ability (Schipolowski et al., 2014). The Visuospatial subtest seemed to be more focused on fluid intelligence. This was seen by its largest statistical associations being with WAIS-IV subtests that index working memory.

Nevertheless, it appears that the Matrix Matching Test is principally measuring the same latent construct as the WAIS-IV, which is taken to be general intelligence (Canivez & Watkins, 2010). In fact, based on the principal component analysis, the Matrix Matching Test was a better measure of this than six of the seven included subtests of the WAIS-IV. However, there are some limitations to the current research. The sample sizes are relatively small compared with the samples often employed in intelligence test validation. Further investigation with the Matrix Matching Test with other samples will allow a clearer idea of its use as a measure of general intelligence. A further issue is its culture dependence;

no intelligence test is completely culture fair, and this is a problem even for tests which rely only on visuospatial reasoning and is likely to be an even bigger problem for the Matrix Matching Test due to the semantic content. Again further research can address this. Without such additional verification, the test is not appropriate for comparison of groups from different cultures.

The Matrix Matching Test presented here is fast to administer—taking on average about 10 minutes. However, there is no time limit and some participants may take longer. The lack of a time limit is deliberate. Several versions of the Progressive Matrices are similarly untimed, and it has been noted that untimed versions tend to be unidimensional in factor analytic studies, while timed versions tend to include an additional factor of processing speed (Raven, Raven, & Court, 1998).

In summary, I have provided a cognitive test that appears to measure general intelligence and appears to have reasonably good psychometric properties. This could be used for academic investigation, providing it is done so from an experimental perspective in which the properties are reevaluated for appropriateness in the specific research contexts.

ORCID iD

Graham Pluck 🄳 http://orcid.org/0000-0002-0368-0051

References

Barbey, A. K., Colom, R., Solomon, J., Krueger, F., Forbes, C., & Grafman, J. (2012). An integrative architecture for general intelligence and executive function revealed by lesion mapping. *Brain*, *135*(4), 1154–1164. doi:10.1093/brain/aws021

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study and Group Behavior*, *1*(1), 23–33.

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*(1), 121–130. doi:10.1037/a0017767

Binet, A. (1903). *L'Etude experimentale de l'intelligence [Experimental study of intelligence]*. Paris, France: Schleicher Freres & Cie.

Bingham, W. C., Burke, H. R., & Murray, S. (1966). Raven's progressive matrices: Construct validity. *Journal of Psychology: Interdisciplinary and Applied*, *62*(2), 205–209. doi:10.1080/00223980.1966.10543785

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): Exploratory and higher order factor analyses. *Psycholological Assessment*, *22*(4), 827–836. doi:10.1037/a0020429

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. doi:10.1037/0033-295X.97.3.404

Cattell, R. B. (1967). The theory of fluid and crystallized general intelligence checked at the 5-6 year-old level. *British Journal of Educational Psychology*, *37*(2), 209–224. doi:10.1111/j.2044-8279.1967.tb01930.x

Cattell, R. B. (1973). *Cattell culture fair intelligence tests*. Champaign, IL: Institute for Personality and Ability Testing.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. doi:10.1016/j.tics.2003.10.005

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201–211. doi:10.1038/nrn2793

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. doi:10.1016/j.tics.2010.01.004

Duncan, J., Chylinski, D., Mitchell, D. J., & Bhandari, A. (2017). Complexity and compositionality in fluid intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(20), 5295–5299. doi:10.1073/pnas.1621147114

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. doi:10.1037/0096-3445.128.3.309

Gottfredson, L. S. (1997). Mainstream science on intelligence (editorial). *Intelligence*, *13*, 13–23. doi:10.1016/S0160-2896(97)90011-8

Hampshire, A., Highfield, R. R., Parkin, B. L., & Owen, A. M. (2012). Fractionating human intelligence. *Neuron*, *76*(6), 1225–1237. doi:10.1016/j.neuron.2012.06.022

Howard, D., & Patterson, K. (1992). *Pyramids and palm trees test*. San Antonio, TX: Psychological Corporation.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to

verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. doi:10.1037/0096-3445.133.2.189

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. doi:10.1016/j.jesp.2013.03.013

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(1), 60–90. doi:10.1037/rev0000039

Meyers, J. E., Zellinger, M. M., Kockler, T., Wagner, M., & Miller, R. M. (2013). A validated seven-subtest short form for the WAIS-IV. *Applied Neuropsychology: Adult*, *20*, 249–256. doi:10.1080/09084282.2012.710180

Park, D. C. (2000). The basic mechanisms accounting for age related decline in cognitive function. In D. C. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 1–21). Philadelphia, PA: Psychology Press.

Pluck, G., Ruales-Chieruzzi, C. B., Paucar-Guerra, E. J., Andrade-Guimaraes, M. V., & Trueba, A. F. (2016). Separate contributions of general intelligence and right prefrontal neurocognitive functions to academic achievement at university level. *Trends in Neurosciences and Education*, *5*(4), 178–185. doi:10.1016/j.tine.2016.07.002

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual section 4: Advanced Progressive Matrices*. Oxford, England: Oxford University Press.

Raven, J. C. (1936). *Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive*. (M.Sc. thesis). London, England: University of London.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. doi:10.1037/a0026838

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition, Technical Manual*. Itasca, IL: Riverside Publishing.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156–158. doi:10.1016/j.intell.2014.05.014

Wechsler, D. (2012). *Escala de Inteligencia de Wechsler Para Adultos-IV [Wechsler Adult Intelligence Scale-IV]*. Madrid, Spain: Pearson.

Author Biography

Graham Pluck, PhD, is a professor of psychology and director of the Quito Brain and Behavior Lab at Universidad San Francisco de Quito, Ecuador, and an honorary research fellow at the University of Sheffield, UK. His research covers various aspects of cognitive and neuropsychology as applied to clinical and social issues.