Internet Cognitive Testing of Large Samples Needed in Genetic Research

Claire M. A. Haworth,[1] Nicole Harlaar,[1] Yulia Kovas,[1] Oliver S. P. Davis,[1] Bonamy R. Oliver,[2] Marianna E. Hayiou-Thomas,[3] Jane Frances,[1] Patricia Busfield,[1] Andrew McMillan,[1] Philip S. Dale,[4] and Robert Plomin[1]

[1] *Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, United Kingdom*
[2] *Forensic Mental Health Science, Institute of Psychiatry, King's College London, London, United Kingdom*
[3] *Department of Psychology, University of York, York, United Kingdom*
[4] *Department of Speech and Hearing Sciences, University of New Mexico, Albuquerque, United States of America*

Quantitative and molecular genetic research requires large samples to provide adequate statistical power, but it is expensive to test large samples in person, especially when the participants are widely distributed geographically. Increasing access to inexpensive and fast Internet connections makes it possible to test large samples efficiently and economically online. Reliability and validity of Internet testing for cognitive ability have not been previously reported; these issues are especially pertinent for testing children. We developed Internet versions of reading, language, mathematics and general cognitive ability tests and investigated their reliability and validity for 10- and 12-year-old children. We tested online more than 2500 pairs of 10-year-old twins and compared their scores to similar internet-based measures administered online to a subsample of the children when they were 12 years old (> 759 pairs). Within 3 months of the online testing at 12 years, we administered standard paper and pencil versions of the reading and mathematics tests in person to 30 children (15 pairs of twins). Scores on Internet-based measures at 10 and 12 years correlated .63 on average across the two years, suggesting substantial stability and high reliability. Correlations of about .80 between Internet measures and in-person testing suggest excellent validity. In addition, the comparison of the internet-based measures to ratings from teachers based on criteria from the UK National Curriculum suggests good concurrent validity for these tests. We conclude that Internet testing can be reliable and valid for collecting cognitive test data on large samples even for children as young as 10 years.

A practical problem in conducting genetic research is that large samples are needed but it is expensive to test large samples in person, especially when the participants are distributed over a wide area. In an attempt to address this problem we have developed cognitive tests that can be administered via the Internet; these tests make it possible to assess large samples efficiently and economically.

The idea of using computers to aid the collection of data is not a new one (see, e.g., Kiesler & Sproull, 1986; Welch & Krantz, 1996). The technological advances in computing and the invention of the World Wide Web in the last 20 years (Abbate, 1999) have made computerized and Internet testing possible on a large scale. Researchers can use the Internet as a tool for recruiting participants as well as administering questionnaires and tests via the Internet. One concern about Internet testing is that the characteristics of Internet samples might differ from samples collected by more traditional methods. In fact, research has shown that Internet samples are reasonably representative in terms of adjustment (Gosling et al., 2004). Internet samples are also generally more diverse in respect to gender, socioeconomic status and age than traditional samples that are often drawn from undergraduate university students (Gosling et al., 2004).

Although there is some consensus that Internet testing is feasible, few studies have considered the validity and reliability of Internet testing, and how results from Internet tests compare to results from paper and pencil tests. All the work in this area to date has focused on questionnaire research, particularly in the field of personality. Findings suggest that, with regard to questionnaire data, the Internet is a valid and reliable tool (see, e.g., Meyerson & Tryon, 2003; Pettit, 2002). An Internet-based neurocognitive screening for adult head injuries has been developed (Erlanger et al., 2002, 2003), but we are unaware of any study that has validated the use of cognitive Internet-based tests in adults or children.

Received 20 February, 2007; accepted 3 May, 2007.

Address for correspondence: Claire Haworth, SGDP Centre, P080, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK. E-mail: Claire.Haworth@iop.kcl.ac.uk

Large Samples for Genetic Research

Quantitative and molecular genetic research requires large samples. This is becoming increasingly pertinent to molecular genetic studies as it becomes accepted that most genetic influence for complex traits involves many quantitative trait loci (QTLs), each with very small effect sizes (Cardon & Bell, 2001), which will require sample sizes in the thousands and possibly tens of thousands to be adequately powered. In-person cognitive testing entails considerable costs arising from employing and training personnel, time and travel expenses incurred by personnel or participants, as well as the costs of data entry. In our experience using UK-wide samples, our average cost per test session in the home is about £170, which quickly becomes untenable with samples in the thousands, particularly for longitudinal research.

Internet tests have three major advantages over traditional methods. Firstly, although moderately costly to prepare, Internet tests can be used repeatedly at no additional expense, which is ideal for testing very large samples (Naglieri et al., 2004). Secondly, scoring is done automatically and immediately, saving time and money and also eliminating data-entry errors (Kraut et al., 2004; Naglieri et al., 2004). Finally, Internet tests provide a method of collecting data on very large samples quickly, and can facilitate the collection of data from diverse samples from all around the world (Naglieri et al., 2004).

The Present Study

Can Internet testing be used as a reliable and valid tool for cognitive testing of children? The present study assessed the reliability and validity of Internet testing of reading, language, mathematics and verbal and nonverbal cognitive abilities in 10- and 12-year-old twins. In addition to examining internal consistency of all measures at both ages, we also investigated 2-year stability from 10 to 12 years. As a direct test of the validity of the 12-year Internet testing, we compared Internet testing of reading and mathematics to similar tests administered in person. Finally, as an index of concurrent validity, we compared Internet testing to teacher reports of achievement in reading and mathematics based on criteria from the UK National Curriculum.

Materials and Methods

Sample

The sampling frame for the present study was the Twins' Early Development Study (TEDS), a study of twins born in England and Wales in 1994, 1995, and 1996 (Oliver & Plomin, 2007; Trouton et al., 2002). The TEDS sample has been shown to be reasonably representative of the general population (Kovas et al., in press a; Oliver & Plomin, 2007).

Internet Access and Connection Speed

In TEDS, 80% of the families have daily access to the Internet (based on a pilot with 100 randomly selected TEDS families), which is similar to the results of market surveys of UK families with adolescents. Most children without access to the Internet at home have access in their schools and local libraries. At 10 years, 73% of those with Internet at home were known to have a broadband Internet connection, and at 12 years this rose to 81%. There was no correlation between Internet speed (i.e., broadband vs. dial-up connection) and socioeconomic status (SES) at 10 or 12 years. The SES measure is a composite of parental education, occupation and the age of mother at the birth of first child.

Estimates of average completion time for the 10-year battery, for children with a broadband connection and those with a dial-up connection, were 82.2 minutes (*SD* = 33.3), and 122.1 minutes (*SD* = 36.2) respectively. At 12 years, average completion time for the whole battery is 97.9 minutes (*SD* = 27.3) for children with a broadband connection, and this is much longer for those children with a dial-up Internet connection (mean = 148.6 minutes, *SD* = 49.0). Due to the length of the 12-year battery, it was separated into two parts, A and B. Information about which tests are in part A and B is included in Table 1. Part B included the language tests, which take longer to download due to the amount of audio streaming required. For this reason, those twins known to have a broadband Internet connection were given part B first and then encouraged to complete part A. Those families with a dial-up Internet connection completed part A first, and then were given access to part B. Again, these families were encouraged to complete both parts of the Internet battery, although they were informed that the download times for the tests in Part B could be quite long on a dial-up connection.

Procedure for Contacting TEDS Families

For the 10-year study the parents of the twins were contacted and asked for consent. Families were then sent log-in packs and information about the Internet battery. A freephone number was available to the families in case of any problems or technical difficulties with the battery. Each family was assigned a caller, who briefly telephoned the family at the beginning of testing, and was able to monitor the twins' progress online. The 10-year study was the first time the TEDS twins had used a test battery on the Internet, and many of the families had technical questions about the Internet and their computers. At 12 years the procedure for contacting the families was the same. Most of the families had already done the Internet testing for the 10-year study and consequently 11% finished the 12-year battery before they were contacted by their caller.

10-Year Sample

From the 1994 and 1995 cohorts of the TEDS sample, 4135 families agreed to participate in the 10-year testing (71%). Of these, 5404 individuals (including 2635 pairs) completed the entire Internet battery (65%). Data were collected for roughly equal numbers of males and females (45% male; 55% female) and of

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

zygosity groups (36% monozygotic [MZ]; 32% dizygotic [DZ] same sex; 32% DZ opposite sex).

12-Year Sample

From the 1994 cohort of the TEDS sample, 1549 families agreed to participate in the 12-year testing (71%). Of these, 1908 individuals (954 pairs) completed part A of the Internet battery (62%) and 1480 individuals (740 pairs) completed part B of the Internet battery (48%); 1138 individuals (569 pairs) completed both parts of the Internet battery. These completion rates are based on all the measures included in each part of the test battery; therefore the N for each individual measure is larger because not all of the children completed all of the tests. The sample was 44% male and included data from all zygosity groups (37% MZ; 33% DZ same sex; 30% DZ opposite sex). The 12-year data only includes children born in 1994; testing of the 1995 and 1996 cohorts is ongoing.

12-Year Subsample

Thirty of the children (15 twin pairs) were reassessed in person within 3 months of completing the Internet-based tests at 12 years. Participants from across the distribution of scores were selected on the basis of the Internet battery.

Measures

In designing our Internet-based battery, it was important that we guarded against potential problems associated with research on the Internet. For example, the testing was administered by a secure server in the TEDS office, which also provided a secure site for data storage; identifying information was kept separately from the data. Appropriate safeguards were in place that prevented children from answering the same item more than once. We provided technical support and assigned a caller to each family, who contacted the family at the start of testing and provided support and encouragement throughout testing. Furthermore, our toll-free telephone number was available to parents and children in case of any problems or questions.

Parents supervised the testing by coming online first with a user name and password for the family, examining a demonstration test and completing a consent form. Then parents allowed each twin to complete the test in turn. Each twin had a unique ID number as well as a family ID number. Parents were urged not to assist the twins with answers and we are confident that most parents complied with this requirement, particularly given that families have been participating in the study for some time, and reliability and validity data for previous measures have been supportive of this assumption (Oliver et al., 2002; Saudino et al., 1998). There is also limited supervision provided by the family caller.

A set of adaptive branching rules was developed, so that all the children started with the same items, but then were branched to easier or harder items depending on their performance (see Kovas et al., in press a, for details). As with many psychological tests that use branching (e.g., Wechsler Intelligence Scale for Children, WISC-III-UK, Wechsler, 1992), the generic scoring rules were as follows: 1 point was recorded for each correct response, for each unadministered item preceding the child's starting point, and for each item skipped through branching to harder items. After a certain number of failures, a discontinue rule was applied within each category, and no points were recorded for all items after discontinuation. As with other psychological tests with items of increasing difficulty and using similar rules, this scoring system is equivalent to that in which all children attempt all items, allowing us to calculate total number and proportion of correct responses for each child, as well as testing the internal consistency of each category. Specific branching and discontinuation rules are available from the authors. Adapting to the children's competence increases their engagement, while limiting the number of items that need to be answered (Birnbaum, 2004). The test battery is self-paced, and can be completed over a period of several weeks. Each child's performance is monitored online and families are telephoned by their caller to provide support and encouragement throughout testing. All measures are normally distributed and show skewness well under one, suggesting that these measures are discriminating at the low and high ends of the distribution.

To create the Internet test battery, we worked with Planet Three Publishing (www.planet3.co.uk) and e-Business Systems (www.e-businesssystems.co.uk).[1]

The costs of our Internet-based testing were largely due to the costs in creating the 10-year battery because we chose to develop our own battery rather than using commercially available products. The cost was much less for the 12-year battery because much of the battery was the same. The production costs of course depend on how much one is willing to invest in creating the battery, especially in terms of graphics that make the tests more interesting to children. Other investments that we chose to make even though they are not intrinsic to Internet-based testing include our callers who intervened if the children had problems with the online testing, and who also monitored the children's progress (Internet-based testing makes this easy to do) and encouraged the children if their progress was flagging. Another expense was that we provided vouchers for the children after they completed the testing, although this is an expense we would have incurred even if we had tested the children in person in their homes. If commercial online tests are available, the costs of the tests are likely to be comparable to the use of traditional paper-and-pencil tests or tests administered by computer. Thus, the difference between Internet testing and in-person testing lies in the time and expense involved in traveling to the children's homes for testing or having them travel to our laboratory. The average cost for in-person testing (£170 in our

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

case) is prohibitive for testing very large samples. In contrast, the production and running costs for Internet-based testing are amortized across the number of subjects tested, which is ideal for testing very large samples.

10-Year Internet Battery

Reading

At age 10, the twins completed an adaptation of the reading comprehension subtest of the Peabody Individual Achievement Test (Markwardt, 1997), which we will refer to as $PIAT_{rc}$. The $PIAT_{rc}$ assesses literal comprehension of sentences. The sentences were presented individually on the computer screen. Children were required to read each sentence and were then shown four pictures. They had to select the picture that best matched the sentence they had read using the mouse. All children started with the same items, but an adaptive algorithm modified item order and test discontinuation depending on the performance of the participant. The Internet-based adaptation of the $PIAT_{rc}$ contained the same practice items, test items and instructions as the original published test. Test–retest reliability of the $PIAT_{rc}$ across 7 months was estimated as .66 in a subsample of 55 twin pairs in TEDS (Harlaar et al., in press).

Mathematics

In order to assess mathematics, we developed an Internet-based battery that included questions from three different components of mathematics. The items were based on the National Foundation for Educational Research 5–14 Mathematics Series, which is linked closely to curriculum requirements in the UK and the English Numeracy Strategy (NferNelson Publishing Co. Ltd, 1999). The presentation of items was streamed, so that items from different categories were mixed, but the data recording and branching were done within each category. The items were drawn from the following three categories: Understanding Number, Non-Numerical Processes and Computation and Knowledge. A composite score was created by calculating the mean of the three mathematics scores. The mathematics battery is described in more detail elsewhere (Kovas et al., in press b).

General Cognitive Ability (*g*)

At age 10, the twins were tested on two verbal tests, WISC-III-PI Multiple Choice Information (General Knowledge) and Vocabulary Multiple Choice subtests (Wechsler, 1992), and two nonverbal reasoning tests, the WISC-III-UK Picture Completion (Wechsler, 1992) and Raven's Standard Progressive Matrices (Raven et al.,1996). We created a *g* score with equal weights for the four tests by summing their standardized scores.

12-Year Internet Battery

The 12-year Internet battery included the same $PIAT_{rc}$ as at 10 years but added another test of reading comprehension and a test of reading fluency. The tests of mathematics were similar although items of greater difficulty were added at 12 years. The same four verbal and nonverbal cognitive tests were used, although a few items of greater difficulty were added to the Raven's measure. Two tests of spatial reasoning were also added. The biggest change at 12 years was the inclusion of three tests of language ability to assess syntax, semantics and pragmatics.

Reading

Three measures of reading ability were used at 12 years: two measures of reading comprehension and a measure of reading fluency.

Reading Comprehension

At 12 years, the same $PIAT_{rc}$ was used to assess reading comprehension. As well as the $PIAT_{rc}$, we assessed reading comprehension at age 12 using the GOAL Formative Assessment in Literacy for Key Stage 3 (GOAL plc, 2002). The GOAL is a test of reading achievement that is linked to the literacy goals for children at Key Stage 3 of the National Curriculum. Questions are grouped into three categories: Assessing Knowledge and Understanding (e.g., identifying information, use of punctuation and syntax), Comprehension (e.g., grasping meaning, predicting consequences), and Evaluation and Analysis (e.g., comparing and discriminating between ideas). Evaluation and analysis is deemed the highest order of the three skills. Within each category, questions about words, sentences, and short paragraphs are asked. Because we were primarily interested in comprehension skills, we used questions from the two relevant categories, Comprehension, and Evaluation and Analysis (20 items from each category). Correct answers were summed to give a total comprehension score.

Reading Fluency

At 12 years, reading fluency was assessed using an adaptation of the Woodcock-Johnson III Reading Fluency Test (Woodcock et al., 2001). This is a measure of reading speed and rate that requires the ability to read and comprehend simple sentences quickly, for example, 'A flower grows in the sky? — Yes/No'. Low performance on reading fluency may be a function of limited basic reading skills or comprehension. The online adaptation consists of 98 yes/no statements; children need to indicate yes or no for each statement, as quickly as possible. There is a time limit of 3 minutes for this test. Correct answers were summed to give a total fluency score.

Language

In order to assess receptive spoken language, standardized tests were selected that would discriminate children with language disability as well as being sensitive to individual differences across the full range of ability. Furthermore, an aspect of language that becomes increasingly important in adolescence — and that shows interesting variability at this age — is

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

metalinguistic ability, that is, knowledge about language itself (Nippold, 1998). For this reason, the three measures selected for testing included one with low metalinguistic demands designed to assess syntax (Listening Grammar) and two with higher demands that assess semantics (Figurative Language) and pragmatics (Making Inferences).

Syntax. Syntax was assessed using the Listening Grammar subtest of the Test of Adolescent and Adult Language (TOAL-3; Hammill et al., 1994). This test requires the child to select two sentences that have nearly the same meaning, out of three options. The sentences are presented orally only.

Semantics. Semantics were assessed using Level 2 of the Figurative Language subtest of Test of Language Competence — Expanded Edition (Wiig et al., 1989), which assesses the interpretation of idioms and metaphors; correct understanding of such nonliteral language requires rich semantic representations. The child hears a sentence orally and chooses one of four answers, presented in both written and oral form.

Pragmatics. Level 2 of the Making Inferences subtest of the Test of Language Competence (Wiig et al., 1989) assessed an aspect of pragmatic language, requiring participants to make permissible inferences on the basis of existing (but incomplete) causal relationships presented in short paragraphs. The child hears the paragraphs orally and chooses two of four responses, presented in both written and oral form.

Mathematics

A revised version of the 10-year mathematics Internet test was administered that followed the same format, but included more advanced questions to reflect the age of the twins. The three mathematics categories were the same as in the 10-year battery (Understanding Number, Non-Numerical Processes and Computation and Knowledge).

General Cognitive Ability (g)

At 12 years the same verbal and nonverbal tests were used (general knowledge, vocabulary, picture completion and Raven's matrices). Raven's matrices test was updated to include more difficult items from the Advanced Progressive Matrices (Raven et al., 1998). We created a *g* score with equal weights for the four tests by summing their standardized scores.

Spatial Reasoning Tests

The spatial reasoning tasks are intended to complement the two nonverbal reasoning tests described earlier in relation to g (these tests are not included in the g score). The Spatial Reasoning series (NferNelson Publishing Co.Ltd, 2002a, 2002b, 2002c) assesses a range of cognitive tasks that involve shape and space, such as mentally combining and rotating shapes, or imagining how a shape would look from different viewpoints. The tests do not require reading, the instructions are very simple, and the tests are not culturally specific. These tests are not timed, but in order for them to be close to the

Table 1

Summary of Internet Measures at 10 and 12 Years

Age at testing	Intended domain	Test	Reference
10 and 12	Reading: Literal comprehension	$PIAT_{rc}$ [A]	Markwardt (1997)
12	Reading: Higher-order comprehension	GOAL formative assessment in literacy [A]	GOAL plc (2002)
12	Reading: Fluency	Reading Fluency [A]	Woodcock et al. (2001)
12	Language: Syntactic	Listening Grammar: Test of Adolescent and Adult Language [B]	Hammill et al. (1994)
12	Language: Semantics	Figurative language: Test of Language Competence [B]	Wiig at el. (1989)
12	Language: Pragmatics	Making inferences: Test of Language Competence [B]	Wiig at el. (1989)
10 and 12	Mathematics	Mathematics: Understanding number; nonnumerical processes; computation and knowledge [A]	NferNelson (2001)
10 and 12	Cognitive: Verbal	General Knowledge [B]	Wechsler (1992)
10 and 12	Cognitive: Verbal	Vocabulary [B]	Wechsler (1992)
10 and 12	Cognitive: Nonverbal	Raven's Standard (and Advanced) Progressive Matrices [B]	Raven et al. (1996, 1998)
10 and 12	Cognitive: Nonverbal	Picture completion [B]	Wechsler (1992)
12	Cognitive: Spatial reasoning	Hidden shapes [AB]	NferNelson (2002a, 2002b, 2002c)
12	Cognitive: Spatial reasoning	Jigsaws [AB]	NferNelson (2002a, 2002b, 2002c)

Note: [A] = test is in part A of the 12-year battery
[B] = test is in part B of the 12-year battery
[AB] = test is included in both part A and B of the 12-year battery

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

Table 2

Internal Consistency Reliability of Web-Based Measures — Cronbach's Alpha Coefficients

Age	Test	Cronbach's alpha	*N*
10 years	Reading: $PIAT_{rc}$	.95	2924
	Mathematics: Understanding number	.92	2595
	Mathematics: Nonnumerical processes	.78	2698
	Mathematics: Computation and knowledge	.93	2698
	Cognitive: Ravens matrices	.91	2614
	Cognitive: Vocabulary	.90	2576
	Cognitive: Picture completion	.74	2569
	Cognitive: General knowledge	.87	2615
12 years	Reading: $PIAT_{rc}$	.94	1069
	Reading: GOAL	.91	1047
	Reading: Fluency	.96	1069
	Language: Syntax	.94	759
	Language: Semantics	.66	846
	Language: Pragmatics	.58	790
	Mathematics: Understanding number	.91	982
	Mathematics: Nonnumerical processes	.88	982
	Mathematics: Computation and knowledge	.94	982
	Cognitive: Ravens matrices	.76	833
	Cognitive: Vocabulary	.88	786
	Cognitive: Picture completion	.72	761
	Cognitive: General knowledge	.81	940
	Cognitive: Hidden shapes	.89	1171
	Cognitive: Jigsaws	.78	1143

Note: This analysis was conducted using one member of each twin pair.

original test format, we encourage children to do them as quickly as they can.

Hidden shapes. Children search for a specific shape embedded in one of four more complex shapes.

Jigsaws. Children decide which shape can be made by combining all four available 'jigsaw' pieces.

Table 1 provides a summary of the Internet measures used at 10 and 12 years and their references.

Tests Administered in Person

Age-appropriate standard versions of the reading comprehension ($PIAT_{rc}$ and GOAL), reading fluency, and mathematics tests were administered using standard test protocol to a subsample of 30 children at the Social, Genetic and Developmental Psychiatry Centre. The standard versions of these tests are much longer than our branched Internet tests, which limit the number of questions each child has to answer. Due to time constraints we were unable to administer the entire cognitive battery. The total testing time was approximately three hours for each participant. Testing was done within 3 months of completion of the 12-year Internet battery (Mean = 65 days, *SD* = 22).

National Curriculum (NC) Measures

As for all UK children at 10 and 12 years, the twins' academic performance was assessed throughout the year by their teachers using the assessment materials of the National Curriculum (NC), the core academic curriculum developed by the Qualifications and Curriculum Authority (QCA). The NC Teacher Assessments consist of teachers giving a score on the basis of the child's performance throughout the school year. Reminders of the NC criteria used to select the appropriate attainment level were provided as part of the questionnaire. Further details about these measures have been published previously (Haworth et al., 2007; Walker et al., 2004). At 10 and 12 years NC measures of reading and mathematics were collected. Of the teacher questionnaires sent, at 10 years, 6129 individual forms (79%) were returned complete, and at 12 years, 2312 individual forms (71%) were returned complete.

Analyses

Internal Consistency Reliability

Cronbach's alpha coefficients were calculated for the Internet-based measures using one randomly selected member of each twin pair. For the mathematics and general cognitive abilities scales, internal consistency reliability was calculated for the items within each subscale.

Bivariate Correlations

Bivariate correlations between pairs of measures were calculated using a phenotypic analysis in the structural equation software Mx (Neale et al., 1999) that controlled for the twin-pair structure of the data, and therefore used the entire dataset. These correlations are equivalent to Generalized Estimating Equation (GEE) correlations. This analysis was similar to that of a recent report (Sluis et al., 2006). Correlations using one randomly selected member of a twin pair produced the same pattern of results. These analyses allowed us to investigate the validity of the Internet-based tests (to what extent are the tests measuring what they were designed to measure) and their reliability (the extent to which the tests are stable). Bivariate correlations were used to investigate three relationships:

Table 3

Stability of Internet Measures of Reading Comprehension, Mathematics and *g* from 10 to 12 Years

Measures	*r* (95% CI)	*N**
Reading Comprehension: 10y $PIAT_{rc}$–12y $PIAT_{rc}$	.57 (.53–.60)	1405
Mathematics: 10y Mathematics–12y Mathematics	.66 (.63–.69)	1206
General Cognitive ability:10y *g*–12y *g*	.66 (.62–.69)	987

Note: $PIAT_{rc}$ is a measure of reading comprehension; *g* = general cognitive ability.
**N* refers to the number of individuals with complete data at both years. The N value for g is lower than for reading and math because it requires children to have complete data for four tests at each age, and in the 12-year battery it is in part B, which children are only advised to complete on a broadband connection.

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

Table 4

Validity: Correlations Between Internet and In-Person Testing

Measures	*r* (95% CI)	*N**
Reading Comprehension ($PIAT_{rc}$)	.80 (.40–.91)	30
Reading Comprehension (GOAL)	.52 (.17–.77)	29
Reading Fluency	.81 (.57–.92)	29
Mathematics	.92 (.83–.97)	30

Note: **N* refers to the number of individuals with complete data

1. The stability of Internet scores in reading comprehension ($PIAT_{rc}$), mathematics and *g* from 10 to 12 years.
2. The relationship between Internet-based measures at 12 years and their equivalent versions administered in person for reading comprehension ($PIAT_{rc}$ and GOAL), reading fluency and mathematics.
3. The relationship between Internet-based measures of reading comprehension ($PIAT_{rc}$) and mathematics versus NC teacher-rated measures of reading and mathematics.

Results

Reliability: Internal Consistency

The internal consistency of the Internet-administered measures was examined, yielding high Cronbach's alpha coefficients, as shown in Table 2. The median of coefficients was .89 (range: .58–.96). Tests administered at 10 and 12 years show similar results for internal consistency.

Stability of Internet Measures of Reading Comprehension, Mathematics and g from 10 to 12 Years

As indicated in Table 3, the 2-year stability of scores on Internet-based tests of reading comprehension, mathematics and *g* are .57, .66 and .66, respectively.

Validity: Correlations Between Internet-Based and In-Person Testing

The correlations between Internet-based scores and scores derived from in-person testing are shown in Table 4 with their 95% confidence intervals. The correlations between these very different testing formats are .80 for $PIAT_{rc}$ Reading Comprehension, .52 for GOAL Reading Comprehension, .81 for Reading Fluency, and .92 for Mathematics. The lower correlation for the GOAL (.52) is discussed later.

Concurrent Validity: Correlations Between Internet-Based Measures and Teacher Ratings of Reading and Mathematics

In order to assess concurrent validity, we correlated children's composite measures of Internet-based performance in reading and mathematics to composite measures of reading and mathematics performance in the classroom as assessed over the school year by their teachers on the National Curriculum Criteria. As indicated in Table 5, the correlation was .42 for reading and .50 for mathematics at 10 years and .45 and .56, respectively, at 12 years. These results suggest considerable concurrent validity given the considerable differences in the content and methods of the Internet-based tests and the teacher ratings.

Discussion

The aim of this study was to investigate the extent to which Internet testing can be used to assess cognitive abilities in children as young as 10 years. Our results show that Internet testing is both a reliable and valid method for collecting such data. In terms of internal consistency, all of the Internet-based measures yielded high Cronbach's alpha coefficients, with a median coefficient of .89. However, internal consistency is not the ideal indicator of reliability for a measure that utilizes a new method (as opposed to new items), because the reliable variance might be, in effect, method variance. In addition, the branching in our study might inflate the internal consistency, because children answer fewer questions when branching is used.

Scores on the Internet-based measures show long-term stability from 10 to 12 years, with an average correlation of .63 for measures of reading ($PIAT_{rc}$), mathematics and *g*. This correlation suggests substantial stability for scores on these measures that were administered 2 years apart, in contrast to conventional test–retest intervals of 2 or 3 weeks.

Most importantly, we directly assessed validity by comparing performance on our Internet-based measures of reading and mathematics to performance on traditional versions of these tests administered in person to the children in our laboratory. The average correlation between performance on the Internet-based and in-person tests was .76, despite the fact that the two test sessions were on average 2 months apart. However, the GOAL test of higher reading comprehension yielded a significantly lower correlation (r = .52) than the tests of reading fluency and mathematics (Cohen, 1988), even though the GOAL has high internal consistency (α = .91) and correlates moderately with other reading Internet tests. The in-person version of the GOAL also correlated

Table 5

Concurrent Validity: Correlations Between Internet Measures and Teacher Ratings of Reading and Mathematics

Measures	*r* (95% CI)	*N**
10y Reading: NC vs. Internet	.42 (.39–.44)	4271
12y Reading: NC vs. Internet	.45 (.40–.49)	1182
10y Mathematics: NC vs. Internet	.50 (.48–.53)	3894
12y Mathematics: NC vs. Internet	.56 (.52–.60)	1088

Note: NC = National Curriculum teacher reports. **N* value is given for individuals with complete data for both the Internet measures and teacher ratings.

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

highly with the other in-person tests of reading and mathematics, and also with the other Internet tests. In fact, the in-person version of the GOAL correlated more highly with the Internet tests of reading than it does with the Internet version of the GOAL, questioning the validity of the Internet-based version of the GOAL.

Finally, correlations between our Internet measures of reading and mathematics and teacher reports averaged .48. This suggests substantial validity because these measurement methods — objective Internet tests at a single brief measurement occasion and teacher ratings based on wide range of performance during the entire academic year — are as different in format as two tests could be. We would not expect these correlations to be as high as those between the direct tests, because the teacher reports encompass many other indicators of performance, such as motivation and interest. The teacher reports are also based on an entire year's performance in stark contrast to the snapshot impression of performance from scores on a single measurement occasion. The moderate correlation between the Internet measures and the teacher reports lends support to the validity of the Internet measures — and this is further confirmed by the comparison with validated in-person tests.

Advantages and Disadvantages of Internet Testing

Disadvantages of Internet testing include the costs of creating such tests, as discussed above. Another limitation is that multiple-choice questions suit the Internet format better than open-ended responses. On the Internet, questions that would normally be read aloud to the child require audio streaming, and this greatly increases the download times for each question. Also the mode of response for in-person testing (e.g., pointing, writing or speaking) may be different from responses collected via computers (e.g., clicking and typing), and it is unknown what differences in test scores, if any, this creates. These factors should be taken into consideration when adapting paper and pencil tests for Internet use. For example, if, during in-person testing, a child would normally give the answer orally, then they should not be penalized for spelling mistakes if they must type the answer for the Internet version. In such a case the scoring procedure must accept possible spellings of the correct response. The difficulties in adapting tests for Internet-based testing are shared with the growing number of tests adapted for computer administration. Tests adapted for computer administration are widely available commercially and tests adapted for the Internet are also increasingly available which will alleviate these problems for researchers in the future.

Market surveys of UK families with adolescents and our own study show that in the UK approximately 80% of adolescents have daily access to Internet at home, and practically all have access to it at schools and libraries. Although the Internet is not readily available to all twins in TEDS, the TEDS Internet sample remains representative of the general population, and is representative of the whole TEDS sample.

Finally, the format of Internet testing results in a lack of supervision and control over the testing environment which may be considered a disadvantage. However, the positive side to this is that the social pressure or embarrassment that might be present in face-to-face testing is reduced (Birnbaum, 2004; Kraut et al., 2004). In our introduction to the Internet testing, we stress that the tests should be taken at a quiet place and time. In any case, the high correlations between our Internet-based measures and supervised in-person testing suggest that children are not cheating during Internet testing, and that the issue of supervision is not important in this sample. Researchers considering the use of Internet testing should bear in mind that TEDS children have been part of the study since infancy and thus may be more motivated to comply with our instructions.

The main advantage of Internet testing is that data from large widely dispersed samples can be collected quickly, cheaply, and as we show here, reliably and validly. Internet-based data collection is less error prone because it does not require human transcription and data entry (Kraut et al., 2004; Naglieri et al., 2004). The medium is well suited to older children, most of whom are competent computer users. It is interactive and enjoyable for children to complete; the test questions are easy to understand with suitable on-screen text, voice instructions, graphics and practice items. Additional games were included in our battery to keep the children interested and to reward them for their efforts. Another major benefit of Internet testing is generally true of computerized testing: adaptive branching. Adaptive branching makes it possible to include a very large pool of items — for example, to assess the extreme low and high ends of the distribution — allowing children to complete a relatively small number of items. With adaptive branching, children are less likely to become bored by having to answer questions that are well below their ability level, or disheartened by having to answer questions well above their ability level.

Conclusions

We conclude that, due to the advances in computing and the widespread availability of high speed Internet connections, it is now possible to use Internet testing to assess cognitive ability in children as young as 10 years. Data gathered via Internet testing is both reliable and valid, and this testing approach is a cheap, quick and efficient method for collecting data on large and diverse samples. Internet testing is therefore a valuable resource for genetic research.

Acknowledgments

We gratefully acknowledge the ongoing contribution of the parents and children in the Twins' Early Development Study (TEDS). TEDS is supported by a

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press

program grant (G0500079) from the UK Medical Research Council; our work on mathematics is supported in part by the US National Institute of Child Health and Human Development and the Office of Special Education and Rehabilitative Services (HD 46167); and our work on school environments is supported in part by the US National Institute of Health (HD 44454). We thank Paul Southcombe of Planet Three Publishing (www.planet3.co.uk) and Sean Heraghty of e-Business Systems (www.e-businesssystems.co.uk) for working with us to create the TEDS Internet-based test battery.

Endnote

1 For further details on the design and use of Internet test batteries, including the TEDS battery, contact Sean Heraghty of e-Business Systems (sean_heraghty@e-businesssystems.co.uk).

Reference List

Abbate, J. (1999). *Inventing the Internet*. Cambridge, Massachusetts: MIT Press.

Birnbaum, M. H. (2004). Human research and data collection via the *Internet*. *Annual Review of Psychology, 55,* 803–832.

Cardon, L. R., & Bell, J. (2001). Association study designs for complex diseases. *Nature Genetics, 2,* 91–99.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Erlanger, D. M., Kaushik, T., Broshek, D., Freeman, J., Feldman, D., & Festa, J. (2002). Development and validation of a web-based screening tool for monitoring cognitive status. *Journal of Head Trauma Rehabilitation, 17,* 458–476.

Erlanger, D., Kaushik, T., Cantu, R., Barth, J. T., Broshek, D. K., Freeman, J. R., & Webbe, F. M. (2003). Symptom-based assessment of the severity of a concussion. *Journal of Neurosurgery, 98,* 477–484.

GOAL plc (2002). *GOAL Formative Assessment: Key Stage 3*. London: Hodder & Stoughton.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about *Internet* questionnaires. *American Psychologist, 59*, 93–104.

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (1994). *Test of Adolescent and Adult Language (TOAL-3)*. Austin, TX: Pro-Ed.

Harlaar, N., Dale, P. S., & Plomin, R. (in press). The ART of reading: Genetic and shared environmental mediation of the association between reading experience and reading achievement in 10-year-old twins. *Journal of Child Psychology and Psychiatry*.

Haworth, C. M. A., Kovas, Y., Petrill, S. A., & Plomin, R. (2007). Developmental Origins of Low Mathematics Performance and Normal Variation in Twins from 7 to 9 Years. *Twin Research and Human Genetics, 10,* 106–117.

Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50,* 402–413.

Kovas, Y., Haworth, C. M. A., Dale, P. S., & Plomin, R. (in press a). The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monographs of the Society for Research in Child Development*.

Kovas, Y., Haworth, C. M. A., Petrill, S. A., & Plomin, R. (in press b). Mathematical ability of 10-year-old boys and girls: Genetic and environmental etiology of normal and low performance. *Journal of Learning Disabilities*.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological Research Online. *American Psychologist, 59,* 105–117.

Markwardt, Jr., F. C. (1997). *Peabody Individual Achievement Test — Revised (Normative Update) Manual*. Circle Pines: American Guidance Service.

Meyerson, P., & Tryon, W. W. (2003). Validating *Internet* research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments and Computers, 35,* 614–620.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59,* 150–162.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. (1999). *Mx: Statistical modeling* (5th ed.). Richmond, VA: Department of Psychiatry, Medical College of Virginia.

NferNelson Publishing Co. Ltd (1999). *Mathematics 5–14 series*. Windsor, UK.

NferNelson Publishing Co. Ltd (2002a). *Spatial Reasoning Age 12 -14*. NFER.

NferNelson Publishing Co. Ltd (2002b). *Spatial Reasoning Age 10 and 11*. NFER.

NferNelson Publishing Co. Ltd (2002c). *Spatial Reasoning Age 8 and 9*. NFER.

Nippold, M. A. (1998). *Later language development: The school-age and adolescent years*. Austin, TX: Pro-Ed.

Oliver, B., Dale, P. S., Saudino, K. J., Petrill, S. A., Pike, A., & Plomin, R. (2002). The validity of a parent-based assessment of cognitive abilities in three-year olds. *Early Child Development and Care, 172,* 337–348.

Oliver, B. R. & Plomin, R. (2007). Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics, 10,* 96–105.

Pettit, F. A. (2002). A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behavior Research Methods, Instruments and Computers, 34,* 50–54.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Advanced Progressive Matrices.* Oxford: Oxford Psychologists Press Ltd.

Raven, J. C., Court, J. H., & Raven, J. (1996). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Oxford: Oxford University Press.

Saudino, K. J., Dale, P. S., Oliver, B., Petrill, S. A., Richardson, V., Rutter, M., Simonoff, E., Stevenson, J., & Plomin, R. (1998). The validity of parent-based assessment of the cognitive abilities of 2-year-olds. *British Journal of Developmental Psychology, 16,* 349–363.

Sluis, S. v. d., Posthuma, D., Dolan, C. V., Geus, E. J. C. d., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence, 34,* 273–289.

Trouton, A., Spinath, F. M., & Plomin, R. (2002). Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood. *Twin Research, 5,* 444–448.

Walker, S. O., Petrill, S. A., Spinath, F. M., & Plomin, R. (2004). Nature, nurture and academic achievement: A twin study of teacher ratings of 7-year-olds. *British Journal of Educational Psychology, 74,* 323–342.

Wechsler, D. (1992). *Wechsler intelligence scale for children — Third Edition UK (WISC-III^{UK}) Manual.* London: The Psychological Corporation.

Welch, N., & Krantz, J. H. (1996). The World-Wide Web as a medium for psychoacoustical demonstrations and experiments: Experience and results. *Behavior Research Methods, Instruments and Computers, 28,* 192–196.

Wiig, E. H., Secord, W., & Sabers, D. (1989). *Test of Language Competence — Expanded Edition.* San Antonio, TX: The Psychological Corporation.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

https://doi.org/10.1375/twin.10.4.554 Published online by Cambridge University Press