Accepted to British Educational Research Journal
Article DOI: 10.1002/berj.3560
Acceptance date: 26 June 2019

How valid are 11-plus tests? Evidence from Kent

Anna Brown [a*] and Sarah Fong [a]

[a] *School of Psychology, University of Kent, Canterbury, UK*

*Corresponding author: Anna Brown, Senior Lecturer in Psychological Methods and Statistics, School of Psychology, University of Kent, Canterbury, Kent CT2 7NP, United Kingdom. E-mail: A.A.Brown@kent.ac.uk

Acknowledgements

The authors are grateful to the leadership and governors of the primary school for providing anonymous test results for this research, and to the administration of the school for preparing anonymised records.

The present research was supported by the University of Kent Social Sciences Faculty Research Fund grant (360 42028) awarded to the first author.

Word count: 7,069

How valid are 11-plus tests? Evidence from Kent

Abstract

Despite profound influence of selection-by-ability on children's educational opportunities, empirical evidence for validity of 11-plus tests is scarce. This study focused on secondary selection in Kent, the largest grammar school area in England. We analysed scores from the 'Kent Test' (the 11-plus test used in Kent), Cognitive Assessment Tests (CAT4), and Key Stage 2 Standardised Assessment Tests (KS2) using longitudinal data of two year cohorts (N1=95, N2=99) from one primary school. All the assessment batteries provided highly overlapping information, with the decisive effect of content area (e.g. verbal versus maths) over task type (e.g. knowledge-loaded versus knowledge-free). Thus, the value in differentiating 'pure' (i.e. knowledge-free) ability in 11-plus testing is questionable. KS2 and Kent Test aggregated scores overlapped very strongly, sharing nearly 80% of variance; moreover, KS2-based eligibility decisions had higher sensitivity than the Kent Test in predicting the actual admissions to grammar schools after Head Teacher Assessment (HTA) appeals have taken place. Finally, the use of multiple pass marks for each Kent Test component as well as the total score was found to increase the chance of false rejection. This study provides preliminary evidence that national examinations could be a good basis for selection to grammar schools; it challenges the use of complex admission rules and multiple decisions and questions the value of 11-plus tests.

Keywords: 11-plus; Kent Test; fluid intelligence; crystallized intelligence

Introduction

Among many controversies surrounding grammar schools, one important concern is the lack of consistency and transparency in selection decisions. This is a pertinent issue in Kent, the largest remaining grammar school area in the country, comprising 35 wholly selective grammar schools and four partially selective schools. To secure a place in one of Kent's grammar schools, the parent first has to register the child to sit the 'Kent Test' (name for 11-plus test used in Kent); the child has to sit the test and either pass the specified score criteria, or, failing that, be put forward by their school for re-consideration by a local Head Teacher Assessment (HTA) panel, or, failing that, enter on appeal (Kent County Council, n.d.). As we can see, the rules are complicated with several decision points, some of them made in private. For instance, HTA panels can override the Kent Test results without the pupil's and parent's knowledge. This lack of transparency is accompanied by the absence of published evidence that this procedure works (selects children who will excel in grammar school).

If children have to be selected[1] on ability as part of the state education system, we must make sure that selection procedures imposed on them are valid, fair and necessary. Unfortunately, not much information is available to the public in relation to any of the above questions, in Kent or the rest of the country. Literature search for empirical evidence pertaining to psychometric properties of 11-plus tests returns single studies from years ago based on small datasets (e.g. Bunting, Saris, & Mccormack, 1987). No validation studies or studies of bias are available from publishers of 11-plus tests. This is surprising given that an established principle of psychometric testing is availability of such information to test users (International Test Commission, 2001). Given the importance of this imposed selection to children's educational prospects, a systematic analysis of psychometric properties of 11-plus tests and selection processes more generally is well overdue. Psychometrically, such analyses must focus on reliability (how precise measurement provided by 11-plus tests is), validity (what 11-plus tests measure and what they predict) and fairness (whether 11-plus tests are biased against any groups). Economically, analyses should include utility (cost effectiveness of the selection procedure).

The present paper has the psychometric focus and aims to contribute empirical evidence of validity of the Kent Test, analysing archival data from two recent cohorts in one primary school. The paper is organised as follows. First, we briefly introduce theory and research important to our conception and analysis of validity. Second, we postulate research objectives and questions, and voice some expectations. Next, we describe our samples, assessments and outcomes available for analyses, and statistical methods we used to analyse them. Next, we describe the results of our analyses, make conclusions and discuss potential implications for policy. Finally, we discuss the limitations of the present study and suggest how they could be overcome in future research.

What does the Kent Test measure?

The original rationale for 11-plus testing back in 1944 was measurement of 'pure' (or 'knowledge-free') ability that, as it was argued, cannot be learned through formal education (Jesson, 2013). Thus, the policy makers assumed that pupils of

[1] The present paper does not discuss whether selection is a good thing – this important question is separate from the question of quality of selection in the selective system. For relevant research, see for example Schagen and Schagen (2003).

different backgrounds would have a fair chance in gaining entry into grammar schools. This rationale is echoed today, as test publishers refer to academic 'potential' (rather than 'knowledge', 'skill' or 'attainment') and claim no 'need for excessive preparation' when advocating the use of 11-plus tests (CEM Centre for Evaluation and Monitoring, n.d.). There are good reasons to be sceptical about these claims. Ample evidence is available that simple practice (retaking cognitive tests) has a large inflationary effect on operational results in high stakes assessments (e.g. Hausknecht, Trevor, & Farr, 2002). More specific to 11-plus, it has been experimentally shown that tutoring improves performance, even if received for as little as three hours (Bunting & Mooney, 2001). This in turn unequally benefits children from affluent backgrounds (Jerrim, 2018).

In an attempt to 'reduce the effect of tutoring' on the Kent Test (Allen, Bartley, & Nye, 2017; BBC News, 2013), in 2014 publisher GL Assessment changed the test composition, replacing two previously 'knowledge-free' components – verbal and numerical reasoning, with two curriculum-aligned components – English and mathematics, thus retaining just one combined reasoning component. The change is puzzling, since it is inconsistent with the previous advocacy for measuring 'pure' abilities as the best deterrent from tutoring. Now that the Kent Test contains both knowledge-loaded and knowledge-free components, there are even more questions about the principles on which the test is built. Thus, Collins (2016) suggested that the test is 'uncertain' as to what it means to measure beyond fulfilling the purpose of selecting the top 30 percent of test takers.

In this paper, we investigate what the Kent Test measures, by empirically examining its relationships with other cognitive measures. Alternative models of human intelligence guided this examination. The *'general intelligence'* model, which goes back to Cattell's single-factor model of 1904, postulates that one general ('*g*') factor is responsible for correlations between all human abilities. This model is useful for many purposes; however, it is usually inadequate to explain differential performance in various ability domains. The *'fluid-crystallized'* model, proposed by Cattell in the 1940s and later developed by Horn, makes the distinction between 'fluid' abilities, which are used for solving novel problems for which prior knowledge or skills are not particularly useful, and 'crystallized' abilities, which are used in tasks requiring consolidated knowledge and skills gained through education. Fluid (or pure) abilities are meant to causally influence the development of crystallized abilities. The fluid-crystallized distinction has been very influential in all domains of psychology, and it has been adopted as the basis for 11-plus testing. Predictions based on the fluid-crystallized paradigm such as the greater influence of genetic component on fluid intelligence, however, have been repeatedly disconfirmed (Johnson & Bouchard, 2005), leading to the development of alternative theories. Such theories identify specialised domains of intelligence relating to the task *content*, such as perceptual speed, spatial, verbal fluency, memory, etc. There is broad consensus in the intelligence literature that common variance in ability domains is underlain by the general mental ability factor 'g' at the apex (Johnson & Bouchard, 2005; Valerius & Sparfeldt, 2014); however the number and content of domains in this hierarchy are often study specific. This is not surprising given that factor analysis extracts common variance from test scores, and the results depend heavily on what measures are in the mix.

In this paper, we will investigate the *construct validity* of the Kent Test as the extent to which it measures psychological constructs in common with other primary assessments, and the extent to which it provides unique information. When 11-plus tests were first introduced, national standardised assessments did not exist. Today, with many cognitive assessments administered routinely in schools, and given very strong

correlations between all of them (Spinath, Spinath, Harlaar, & Plomin, 2006), it is doubtful that yet another test can provide fundamentally new information on children attainment, compared, for instance, to Key Stage 2 exams and teacher assessments. Is the additional stress placed on already over-assessed young children (McDonald, 2001) worthwhile, and are the taxpayer moneys well spent on administering this particular assessment?

How accurate is the Kent Test in classifying passes and fails?

To pass the Kent Test, a pupil must achieve the total score of 320 across three scored components (English, mathematics and reasoning) as well as a minimum of 106 in each scored component. This *combined* criterion selects only all-round high scorers, and is referred to as *non-compensatory* approach (the name reflects the lack of opportunity to compensate for a lower score on one component with a higher score on another). Why is this particular approach taken? We could not find a clear answer or proven evidence base; interestingly, another selective county, Buckinghamshire, uses the total score or *compensatory* approach (Allen et al., 2017).

Any eligibility criteria will be subject to errors of classification inherent to the test scores on which they are based. Test theory dictates that on psychological attributes that cannot be directly observed (such as aptitude) but inferred from multiple indicators (such as test items), children with the 'true' score exactly at the pass mark get misclassified 50% of the time. This is true for any test; however, how quickly the classification errors subside as the true score moves away from the pass mark depends on the Standard Error of measurement (SEm) of the test. Unfortunately, the 11-plus test publishers do not make available information on classification accuracy, nor do they publish the SEm for different score levels necessary to calculate this. The misclassification questions are further complicated by the use of the combined criteria because each of the four yes/no decisions (one per each test component as well as the aggregated score) is open to misclassification errors.

In this paper, we will attempt to assess the classification accuracy of the Kent Test by assuming a best-case scenario (small) SEm and simulating samples where the observed scores have the same characteristics as our cohorts, but the 'true' and 'error' components of the scores are known. This will allow us to estimate what percentage of children in our cohorts could have been misclassified, for instance passed (have observed score over the pass mark) when they should have failed (have true score below pass mark) and vice versa. We will examine in this way the current combined eligibility criteria as well as its simplest alternative – a total score criterion.

What role does the Kent Test play in admission decisions?

The role of the Kent Test in selection decisions is not straightforward because the specified score criteria is not the only determinant of the admission decision. For instance, the Kent County Council allows HTA panels to overturn the Kent Test-based eligibility decisions. This procedure can have advantages and disadvantages. Correcting obvious false-negative decisions for children who are known good performers but failed on the day due to irrelevant situational factors (e.g. anxiety, illness) would certainly be advantageous, while subjectivity introduced at this late selection stage could be among disadvantages. Furthermore, the eligibility decisions can be further challenged through the appeals procedure.

In this paper, we will investigate the role that the Kent Test plays in selection decisions by empirically examining the overlap between test-based eligibility decisions with actual admission decisions after the HTA panels have taken place (but before any individual appeals). We will again evaluate the current combined eligibility criteria against a total score criterion based on both Kent Test and Key Stage 2 examinations. These investigations pertain to *criterion-related* validity of the Kent Test scores, if we consider admission decisions as the outcome of selection (criterion). Of course, other criteria can be of interest, with perhaps the most important being future (secondary) academic performance, but we cannot address this question with the data we have, and leave it to future research.

Objectives and research questions

The aim of the present study is to examine validity of the Kent Test scores by analysing them against other primary assessments, and against admission decisions for two cohorts of pupils from one state primary school in Kent. The first objective was to examine *construct validity* of Kent Test scores; in particular, to investigate what they capture in common or in addition to the national curriculum exams. To this end, we investigated attribution of variance in 10 tests from 3 assessment batteries – Cognitive Assessment Tests or CAT4, Key Stage 2 National Curriculum Assessments or KS2, and Kent Tests. We mapped each test according to its content area (i.e. verbal, numerical, and figural) and its contribution of learning (i.e. 'knowledge-loaded' and 'knowledge-free'), and examined the overlap between similar and dissimilar types of tests. If the assessments with different contents but similar contribution of learning (for example, Kent Test mathematics and Kent Test English) correlate at least to the same extent as the assessments with similar content but different contribution of learning (for example, Kent Test mathematics and CAT4 quantitative), then the role of knowledge (and therefore the fluid-crystallized paradigm) is decisive, and the unique contribution of 'fluid' abilities, and therefore the potential value of the knowledge-free component of Kent Test is supported. The opposite would support the primary role of content area and refute the added value of the knowledge-free component. The same hypotheses were tested more formally, by comparing alternative ability models using Confirmatory Factor Analysis (CFA). The importance of learning, for example, would be supported by a model in which a 'fluid' factor is needed to explain variance in Kent Test reasoning component over and above any factors underlying KS2 assessments.

The second objective of this study was to examine *classification accuracy* of the Kent Test. How frequently would the eligibility decisions based on the observed Kent Test scores, using the current combined rules with multiple pass marks, correspond to decisions made on the basis of true scores? And what would the classification accuracy be under a simpler total score criterion?

The third objective was to examine *criterion-related validity* of Kent Test scores, with the criterion being admission decisions once Head Teacher Assessments have taken place. How accurate (sensitive and specific) is the Kent Test in predicting the admission decisions? Furthermore, could the KS2 exam results be used as an alternative basis for selection? The HTA panels are supposed to correct for obvious 'false negatives' – children who have performed consistently well throughout the primary years but failed the 11-plus exams. As the final exams of primary learning, KS2 are likely to contain valuable information on sustained academic performance, and therefore may predict the actual selection decisions with accuracy similar to the Kent Test.

Methods

Sample

We consider two recent[2] cohorts from a state primary school in Kent. The school's admission policy gives priority to children who live close to the school, so the pupils are broadly representative of the local population, with a good range of social, educational and religious backgrounds. Cohort 1 comprised $N_1 = 95$ children (49 boys and 46 girls). Cohort 2 comprised $N_2 = 99$ children (58 boys and 41 girls). The cohorts consisted of entire year populations, except one or two children in each cohort who moved in or out of school after Year 6 began, and therefore had either KS2 results or results of eligibility assessments for grammar education unavailable.

Measures (assessments)

Anonymised results on the following assessments were considered.

Cognitive Assessment Tests (4th edition or CAT4) is a battery of tests assessing reasoning abilities that some schools choose to administer towards the end of Year 5 as an early indicator of progress toward 11-plus tests and KS2 exams approaching in Year 6. CAT4 includes four multiple-choice tests – verbal, quantitative, non-verbal and spatial reasoning. Verbal reasoning includes classification and analogies tasks; quantitative (or numerical) reasoning includes number analogies and number series; non-verbal reasoning includes figure classification and figure matrices; and spatial reasoning includes figure analysis and figure recognition. CAT4 scores are Standard Age Scores (SAS) corrected for pupil's age in days by the test publisher GL Assessment (2008), and standardised nationally to achieve a scale[3] with mean 100 (national average) and standard deviation 15.

Kent Test is the name for 11-plus examinations published by GL Assessment and administered by Kent County Council. The test is voluntary – parents have to pre-register their children in July of Year 5 to sit the test in September of Year 6. Since 2014, the Kent Test has comprised four assessments – English, mathematics, reasoning and creative writing; only the first three multiple-choice components are scored whereas the creative writing paper may be considered in appeals (Kent County Council, 2018). The English assessment includes comprehension, spelling, grammar and punctuation tasks. The mathematics assessment includes a variety of topics that able pupils typically master by the beginning of Year 6. The reasoning assessment is the only remaining component designed to assess 'knowledge-free' ability, and includes verbal, non-verbal and spatial reasoning tasks, similar to those covered by CAT4. The Kent Test scores are age-adjusted and standardised on the population of applicants to Kent grammar schools.

Key Stage 2 National Curriculum Assessments (KS2) are administered to all pupils at the end of Year 6. In the recent years, KS2 have comprised English reading, English grammar and mathematics assessments, assessing knowledge and skills based on the national curriculum. The raw scores are scaled to range between the minimum of 80 and the maximum of 120 with a score of 100 indicating 'the pupil has met the expected standard in the test' (Standards & Testing Agency, 2016). The scaling allows

[2] Exact years are not given to protect privacy of the school and pupils

[3] In psychometric literature, this scale is called 'Deviation IQ'.

comparison over time as the difficulty might vary from year to year. Unlike in the CAT4 or Kent Test batteries, **no** adjustment for age is applied in KS2.

Outcomes (admission decisions)

Two outcomes of selection to grammar schools are considered in this study.

Eligibility assessment is based solely on the results of Kent Test and has two possible outcomes – eligible for Grammar, or for High school (G or H respectively). To be eligible for a grammar school place, a pupil must achieve the total score of 320 across the three scored components of the Kent Test, as well as a minimum of 106 in each component. Those pupils who did not sit the Kent Test are **not** eligible for grammar school and therefore are automatically assigned for high school.

Admission decisions (G or H) are reached after Head Teacher Assessment (HTA) panels have taken place. The procedure is an opportunity for the primary school to appeal against failed eligibility assessments for pupils who are deemed suitable for grammar education, by presenting the pupil's recent assessments and class work for consideration by a panel. Approximately 20% of grammar school places in Kent are awarded through successful HTA appeals (Allen et al., 2017).

Statistical Analyses

To control for possible year-to-year variations in content or difficulty of the assessments, we analyse the data and report results separately for each year cohort.

Missing data

Since CAT and KS2 assessments are administered to all children, any missing results are due to child's absence from school on the day, which can be assumed a random process not affecting the distribution of scores in any systematic way. On the contrary, children (with input from their families and often teachers) self-select to take the Kent Test, and these decisions are commonly informed by past performance on various assessments including CAT4. Therefore, the Kent Test score distribution is affected in a systematic way by missing data, with the present scores tending to be higher scores. This *restriction of range* will typically bring the score mean up, the variance down, and will attenuate (bring down) correlations with other measures (Wiberg & Sundström, 2009). To obtain a more complete picture, as if all the pupils had taken the Kent Test, maximum likelihood (ML) estimation is performed. This analysis assumes the normal distribution of scores in the population, and estimates the sample statistics for those with and without missing data conditioning on all the present measures. We use ML for estimating the unrestricted means and covariance structure, and the restricted confirmatory factor analyses described below.

Analysis of assessment scores correlations

We first examine ML estimated correlations of all assessment scores by cohort. To illustrate our hypotheses concerning attribution of variance, correlations in Table 2 are blocked, shaded or bolded. Correlations between assessment within the same battery or using the same *method* (CAT4, Kent Test or KS2), are blocked on the diagonal. Correlations between assessments that require similar contribution of learning, or using the same question *format* (knowledge-free/fluid or knowledge-loaded/crystallized) are shaded. All assessments within CAT4 are considered fluid, and within KS2 crystallized; while the Kent Test assessments are mixed, with English and mathematics crystallized

and reasoning fluid. Correlations between assessments that involve similar *content* (for example, mathematics) are bolded. In psychometrics, these are called *convergent* correlations. When mapping the tests to content domains, we adopted a simple classification into verbal, numerical and figural abilities (Valerius & Sparfeldt, 2014), which had a good conceptual fit with the measures in this study. Finally, the remaining cells – not blocked, shaded or bolded – represent the *discriminant* correlations, indicating the extent to which the assessment scores overlap when neither content, nor method, nor test format is the same.

To evaluate construct validity of the Kent Test, we focus on correlations involving Kent Test scores only (in the middle block of rows and the middle block of columns). Only for those correlations involving Kent Test components, we will compute averages – average convergent, average discriminant, average method etc.

Confirmatory Factor Analysis of assessment scores

To examine the Kent Test's construct validity more formally, we used confirmatory factor analysis (CFA). Four alternative CFA models were tested as follows.

1) *General intelligence model*, where all tests are underlain by one '*g'* factor;
2) *Fluid-crystallized model*, where tests are underlain by two correlated factors – fluid or crystallized intelligence, depending on contribution of *learning* to performance (i.e. CAT are fluid; KS2 are crystallised; Kent Test are mixed);
3) *Verbal-numerical/figural model*, where subtests are underlain by two correlated factors –verbal or numerical/figural, according to tested *content* (i.e. CAT verbal, KS2 English reading and grammar, and Kent Test English are verbal, and the rest are either numerical or figural, including all mathematics, nonverbal and spatial tasks).
4) *Verbal-numerical-figural model*, where the content mapping is more specific than in the above model and the numerical domain is separated from figural. Thus, there are three correlated factors – verbal (all verbal reasoning and English assessments), numerical (all quantitative / numerical / maths assessments) and figural (all nonverbal, spatial and reasoning assessments).

The fluid-crystallized model signifies the importance of learning; and would support the Kent Test uniqueness in capturing knowledge-free reasoning abilities compared to KS2. The verbal-numerical-figural split, on the other hand, would refute the unique contribution of Kent Test compared to KS2 because both batteries would be underlain by the same content-based factors. The general intelligence model would refute uniqueness of any content or format, since it would assume that variability in all tests are due to one factor.

We fitted each of the CFA models to each cohort separately. To evaluate exact fit of models to data, we considered the chi-square statistic (with significant results indicating the lack of exact fit) and the Standardised Root Mean square Residual or SRMR – a direct measure of discrepancy between observed and model-predicted correlations – with values under 0.08 indicating close fit (Hu & Bentler, 1999). We also considered Comparative Fit Index (CFI) with values over 0.95 indicating close fit, and Root Mean Square of Approximation (RMSEA) with values under 0.06 indicating close fit (Hu & Bentler, 1999). Since some of the examined models are not nested within each other (some are, for example, all the models are nested within the general intelligence model), we also consider the Bayesian Information Criterion (BIC) to enable direct model comparison. BIC heavily favours more parsimonious models, and a model with

the smallest BIC should be preferred as providing the best balance between fit and parsimony (Burnham & Anderson, 2004).

Analyses of eligibility decisions

To estimate classification accuracy of the Kent Test in our cohorts, we carried out a simulation study[4]. Because information on the SEm around the Kent Test pass mark is not available, and we do not have item-level data to estimate this ourselves, we resorted to using the figures[5] for CAT4 tests that are published by the same company and are available (GL assessment, 2012). The reliabilities are reported to range between .87 and .89 for the four individual CAT components. We assumed the best case scenario – reliability of .90 for every Kent Test component, which, by definition, corresponds to 90% of the observed score variance being due to true score. We simulated true and error scores for each test so that they independently accounted for 90% and 10% of the observed score variance, respectively, and so that the observed scores (sum of true and error) were distributed with means, variances and covariances exactly as seen in Cohorts 1 and 2 (see Table 1). To capture a whole range of scores and random variations, we simulated 1000 samples of 1000 hypothetical children.

We then summed the 'true' and 'error' score components that were generated for every hypothetical 'child' to produce the 'observed' scores for English, Mathematics and Reasoning, and summed these to produce the total score. We calculated the 'observed' eligibility based on the current *combined* criterion, with the outcome 1 (pass) if the total score was no less than 320 and the single test scores were all no less than 106, and outcome 0 (fail) otherwise. We calculated the 'true' eligibility in the same way, but using only the generated 'true' score components. Finally, we identified the number of classification errors by counting cases with observed passes but true fails (false positive eligibility decisions) and observed fails but true passes (false negative eligibility decisions) in each replicated sample, and averaged these figures. Using the same steps, we tested an alternative *aggregated* criterion, which imposes a cut-off on the total score only.

Analyses of admission decisions

To assess the role of the Kent Test in admission decisions, we cross-tabulated the eligibility decisions based on the Kent Test score combined criteria with the actual admission decisions after the HTA panels. Overall *accuracy* of the eligibility decisions (and therefore the Kent Test criterion-related validity) was assessed as the percentage of correctly predicted (G-G and H-H) admission decisions. *Sensitivity* (ability to correctly identify all those admitted) was computed as the ratio of true positive (eligible=G and admitted=G) decisions to all positive (admitted=G) decisions. *Specificity* (ability to correctly identify all those not admitted) was computed as the ratio of true negative (eligible=H and admitted=H) decisions to all negative (admitted=H) decisions.

Again, we tested both the current *combined* criterion and the alternative *aggregated* criterion. We also applied the two eligibility criteria to both the Kent Test and the KS2 scores as follows.

1) *Combined.* Aggregated KS2 score no less than 315 and no single test score less than 104 for G = grammar school decision; otherwise H = high school decision. This criterion mirrors directly the current combined Kent Test criterion, with the slightly

[4] Mplus syntax for this simulation study is available from the first author on request.

[5] Only one reliability/SEm figure per test is available; this classical test theory treatment assumes that every ability level is measured with the same precision. This is rarely the case, and ideally estimates for each score level should be derived using Item Response Theory.

lower cut-offs imposed on KS2 (104 instead of 106, and 315 instead of 320) because of the different KS2 scores metric, with lower range and SD (see Table 1). The KS2 cut-offs approximately correspond to the Kent Test cut-offs in terms of the number of standard deviations from the mean.

2) *Aggregated.* Aggregated KS2 score no less than 315 for G = grammar school decision; otherwise H = high school decision. This type of criterion can also be applied to Kent Test, by adopting the cut-off 320 for aggregated Kent Test score.

Results

Descriptive statistics

Table 1 presents the descriptive statistics of the test scores by cohort. The main results relate to observed sample statistics based on the actual numbers of pupils taking the assessments. Only parts of the cohorts took the Kent test ($N_1 = 69$ and $N_2 = 75$), resulting in notable restriction of range for the Kent Test scores as discussed in 'Missing Data' section. While the observed means for CAT4 and KS2 are around 104 (slightly higher than the national average), the means for Kent Test are substantially higher at around 110. However, the ML estimation projected the statistics for the whole sample to be in line with the other batteries (see values in parentheses in Table 1).

TABLE 1 NEAR HERE

Convergent and discriminant validity of the Kent Test

Table 2 provides ML estimated[6] correlations between all assessments by cohort. All the correlations are positive and large, averaging at .67 for Cohort 1 and .77 for Cohort 2. Correlations between aggregate battery scores (not in the table) are very strong – Kent Test aggregated score correlated with CAT4 aggregated score at .84 and .89 for Cohorts 1 and 2 respectively; and with KS2 aggregated score at .88 and .89 for Cohorts 1 and 2 respectively. The latter result is important to note – as KS2 aggregate score explains 77% and 79% of variance in the Kent Test score for Cohorts 1 and 2, it can be considered as a potential alternative to Kent Test. We will examine how this very similar score fares in predicting admission decisions in the last section of Results.

The average correlation of Kent Tests[7] with other tests assessing similar content (*convergent*, in bolded cells) is .72 for Cohort 1 and .80 for Cohort 2. The average correlation of Kent Tests with tests assuming similar contribution of learning (using the same test *format*, in shaded cells) is .68 for Cohort 1 and .78 for Cohort 2. The average Kent Test within-battery correlation (*method*–related, blocked on the diagonal) is .64 for Cohort 1 and .82 for Cohort 2. Finally, the average *discriminant* correlation (all other cells pertaining to Kent Test) is .64 for Cohort 1 and .76 for Cohort 2. For both cohorts, therefore, content-related (convergent) correlations are slightly stronger than the format-related correlations. This provides support for the primary role of content over contribution of learning; however, the magnitude of all types of correlations is very similar, necessitating a more formal analysis of factorial structure.

---

[6] The actual correlations based on available N for each assessment are given in the Supplement (table S1), showing a notable attenuation for all Kent Test correlations as expected.

[7] Note that only correlations pertaining to Kent Test are averaged in these analyses

---

TABLE 2 NEAR HERE

Constructs measured by the Kent Test

For both cohorts, one factor explained the vast majority of variance in test scores (the first and second eigenvalues were 7.05 / 0.83 for Cohort 1 and 7.94 / 0.44 for Cohort 2). Table 3 summarises goodness of fit for all tested CFA models, and Table 4 provides standardized factor loadings for the respective models. Table 4 can also be used as reference for the mapping of particular tests to fluid/crystallized or content factors.

TABLE 3 NEAR HERE

The *general intelligence ('g') model* was inadequate for the Cohort 1 data according to all fit indices except SRMR, which indicated acceptable fit; however, it fitted reasonably well the Cohort 2 data according to CFI and SRMR (but not RMSEA).

The *fluid-crystallized model* was not much better than the nested 'g' model, with negligible improvements in all fit indices. It still could not be considered a close-fitting model for Cohort 1; however, it fitted reasonably well to Cohort 2 data, at least according to CFI and SRMR (but not RMSEA). The fluid and crystallised factors correlated very strongly, at .956 for Cohort 1 and .983 for Cohort 2, indicating the lack of discriminant validity for these constructs.

The *verbal-numerical/figural model* fitted the data from both cohorts well according to CFI and SRMR; and for Cohort 2, the fit was excellent according to all indices including chi-square, which was insignificant indicating exact fit. The model provided a substantial improvement over the nested 'g' model according to all indices. The verbal and numerical/figural factors correlated weaker (.878) than the fluid and crystallized factors for Cohort 1, indicating more discriminant validity for content-based constructs; however, the correlation was still very strong (.958) for Cohort 2.

The *verbal-numerical-figural model* fitted the data similarly well to the nested verbal-numerical/figural model, with all fit indices showing only trivial improvements from the addition of another factor. For Cohort 1, the verbal factor correlated with figural at .904 and with numerical at .849 and figural correlated with numerical at .983. For Cohort 2, the three correlations were uniformly strong – verbal/figural .935, verbal/numerical .954 and figural/numerical .959.

Comparing all the models, nested or not, BIC was the smallest for the 2-factor verbal-numerical/figural model, followed closely by the 3-factor verbal-numerical-figural model. Out of the remaining models, BIC was indecisive between the 'g' and the fluid-crystallized model, favouring the former for Cohort 2 but the latter for Cohort 1; however, the BIC differences for the respective models were very small. Overall, the CFA results are decisive about the primary role of content rather than test format (contribution of learning) in assessments.

TABLE 4 NEAR HERE

Classification accuracy of the Kent Test

The simulation study estimated that under the current *combined* criterion, in a population of Kent Test takers with the score distributions as in Cohort 1, 10% of cases would have been misclassified, with 4.0% falsely passed and 6.0% falsely failed. In Cohort 2, the same total of 10% of cases would have been misclassified, but with a larger imbalance of 3.2% falsely passed while 6.8% falsely failed. False rejections, therefore, were estimated to be more prevalent than false admissions in either cohort.

If the *aggregated* criterion were used, in Cohort 1 the total of 6.8% of cases would be misclassified, with 3.3% false positive and 3.5% false negative decisions. In Cohort 2, 6.2% would be misclassified, with 3.0% false positive and 3.2% false negative decisions. The use of the aggregated criterion, therefore, would improve the classification accuracy by reducing the prevalence of false rejections to the level of false admissions.

Role of the Kent Test in admission decisions

Cross-tabulations of the eligibility decisions based on Kent Test and KS2 and the actual admission decisions after HTA panels are presented in Table 5 (for the current combined criteria) and Table 6 (for the alternative aggregated criteria).

The *combined* KS2 eligibility decisions agreed with the admission decisions in 81.1% of cases for Cohort 1 and in 83.8% of cases in Cohort 2 (see Table 5). Despite slightly weaker overall prediction accuracy than that of the Kent Test combined criterion (just over 85% for both cohorts, boosted by the 100% *specificity* by design[8]), the KS2-based classification yielded higher *sensitivity* (ability to correctly identify students who were actually admitted by HTA panels). The lower sensitivity of the Kent Test combined criteria was due to more 'false negatives', judging unsuitable 19.4% and 23.3% of children in Cohorts 1 and 2 respectively, while they were subsequently judged eligible by the HTA panels. This suggests that KS2 exams correctly predicted some of the manual corrections resulting from the HTA procedure.

TABLES 5 AND 6 NEAR HERE

The *aggregated* KS2 criteria agreed with the admission decisions in 77.9% of cases for Cohort1 and in 88.9% of cases in Cohort 2. Despite the further improved performance in sensitivity compared to KS2 combined criterion, particularly for Cohort 2, where KS2 aggregated criteria with sensitivity 92.5% outperformed the combined criteria based either on Kent Test or KS2, the winner here was the Kent Test. If the aggregated Kent Test criteria were applied to selection in the focal primary school, instead of the current combined criteria, almost all admission decisions would be predicted correctly (93.7% for Cohort 1 and 98.9% for Cohort 2). This suggests the implicit salience of the aggregated Kent Test score (and therefore of the compensatory model) in HTA panel decisions, despite the declared policy of relying on the non-compensatory model.

[8] Once the Kent Test combined criterion is fulfilled, eligibility for grammar cannot be denied. This results in 100% specificity, or ability to correctly identify all those not admitted, through having 0% false positive outcomes.

Discussion

The present study began answering some important questions about 11-plus tests, by examining the Kent Test used in the largest remaining grammar school area in the country. Anonymous archival data from two recent cohorts of pupils, namely scores on three assessment batteries taken longitudinally – CAT4, Kent Test, and KS2 exams – were obtained from one primary school in Kent. The Kent Test scores were examined with respect to variance they shared with other assessments (construct validity), and with respect to agreement with admission decisions resulting from the Head Teacher Assessment panels (criterion-related validity).

The analysis of attribution of variance suggested that all assessments provide highly overlapping information. Ordering of children on all tests concurrently and over time is remarkably consistent, with one general factor explaining most variability in test scores. Beyond the general overlap, convergent correlations between tests measuring similar content (e.g. verbal versus numerical) are slightly stronger than correlations between tests using similar format (knowledge-free versus knowledge-loaded). This provides support for the primary role of content over contribution of learning, and questions the traditional advocacy in 11-plus testing for measuring 'pure' abilities as precursors of all other achievements.

It remains to be seen what 11-plus components best predict future academic performance – in grammar or high schools, and beyond; however, the present study does not allow this examination. Perhaps it is the past performance (i.e. learned knowledge and skill together with the motivation and effort that went into developing those) that should receive more attention in predicting future performance? If so, we have plenty of information on children's performance during their primary school years, including teacher assessments, in-class assessments, and national curriculum assessments. It is doubtful that one relatively short examination paper can provide more reliable information than all such longitudinal information taken together.

The present study suggested that the KS2 exams could be a suitable alternative to 11-plus tests (or be part of an alternative selection system), demonstrating not only good coverage of the latent constructs measured by the Kent Test but also resulting in similar selection decisions. Interestingly, KS2 results aligned more closely with the admissions resulting from the Head Teacher appeals than the Kent Test results. What is it that the curriculum-based school examinations and the HTA panels capture in common over and above 11-plus tests? Perhaps, again, it is the sustained performance, assessed explicitly by the curriculum-based school examinations that matters to the Head Teachers' implicit judgements?Importantly, the analyses also suggested that the use of a single cut-off on the aggregated score rather than multiple cut-offs (one per each test component as well as the aggregated score) aligns better with HTA judgements, who seem to be implicitly reliant on the overall Kent Test score when overturning negative eligibility decisions. The data at hand suggest that if the aggregated criterion were used instead of the current combined criteria, there would be no need for interventions from the HTA panels, at least in the focal school. The compensatory model (one cut-off on the aggregated score) could be more appropriate for finding and promoting children with outstanding talents in at least one area, even if they lack aptitude in others. It can be argued that excluding such children with the current non-compensatory model denies them access to best conditions for developing their particular talents, which may be especially problematic in terms of social exclusion of children who lack the consistency of education due to social or economic disadvantages.

Combined or aggregated, eligibility criteria must minimise the impact of classification errors that are inevitable with the use of any tests and examinations. Our analysis of classification accuracy suggests that the combined eligibility rule, which passes children only when all test components and the total mark are passed but fails them when only one mark is failed, is actually negatively biased – that is, it propagates false rejections. We believe that these results are robust and will likely hold for any tests with similar reliabilities and correlations among the components, for example KS2 assessments. It is the combined criterion itself that capitalises on chance of failing, rejecting with the accuracy of its least reliable component, but admitting with a much greater accuracy by using all available information. We are therefore concerned with its use in practice.

Policy implications

Although no policy changes can be advocated on the basis of the present study, limited to one school only, further investigations are certainly warranted. The following questions, clearly relevant to policy, need to be answered:

1. What is the rationale, except convenient timing in the grammar admission process, for commissioning and administering a standalone entry exam, when plenty of reliable information is already available on children's aptitude and attainment throughout primary years?
2. What is the benefit of using combined eligibility criteria, if most negative eligibility decisions get overturned by HTA panels in favour of evidence suggested by the single aggregated score?
3. What is the value of Head Teacher Assessment panels, if their decisions are so influenced by the aggregated Kent Test score, and also by evidence of sustained performance that is captured by the national curriculum exams?

Limitations and future research

The obvious limitation of this research is its small scale. Only one primary school participated, although two cohorts were examined for replicability. On the positive, the school is large and representative of the area, with a good range of backgrounds. Future research should attempt analysis of Kent Test data on a larger scale. Unfortunately, getting access to suitable data proves very difficult. Although the Kent County Council has recently released Kent Test scores and outcome decisions for one year in response to a freedom-of-information request, these anonymous records cannot answer the questions of the present research since they do not include pupils who did not sit the Kent Test, and cannot be matched to any other attainment scores. Furthermore, no information relevant to the Kent Test results or outcomes are recorded in the National Pupil Database (where longitudinal data about demographics and attainment of pupils in the UK are kept). Addressing these challenges would be an immense step forward.

The second major limitation is that we could not examine fairness of the Kent Test in terms of its social inclusion because no indicators of socio-economic status (for example, free school meals) were available to us. This is due to the obligation by the school to protect this sensitive information, particularly when the overall samples are small. In any effort to obtain larger datasets for analyses in the future, socio-economic status must be made part of the analyses, to see whether selection criteria are fair to children from all backgrounds.

References

Allen, R., Bartley, J., & Nye, P. (2017). *The 11-plus is a loaded dice. Analysis of Kent 11-plus data*. London: Education Datalab. Retrieved from https://educationdatalab.org.uk/wp-content/uploads/2017/05/The-11-plus-is-a-loaded-dice-Report.pdf

BBC News. (2013). New Kent 11-plus makes tutoring "less effective." *BBC News*. Retrieved from https://www.bbc.co.uk/news/uk-england-kent-24668503

Bunting, B., & Mooney, E. (2001). The Effects of Practice and Coaching on Test Results for Educational Selection at Eleven Years of Age. *Educational Psychology*, *21*(3), 243–253. https://doi.org/10.1080/01443410120065450

Bunting, B., Saris, W. E., & Mccormack, J. (1987). A Second-order Factor Analysis of the Reliability and Validity of the 11 plus Examination in Northern Ireland. *The Economic and Social Review*, *18*(3), 137–147.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

CEM Centre for Evaluation and Monitoring. (n.d.). CEM Select. Entrance Assessments. Retrieved October 5, 2018, from https://www.cem.org/entrance-assessments

Collins, M. (2016). On the history of "IQ" and aptitude testing – with specific relation to the Kent Test. Retrieved October 4, 2018, from http://kenteducationnetwork.org/2016/06/on-the-history-of-iq-and-aptitude-testing-with-specific-relation-to-the-kent-test/

GL assessment. (2012). *CAT4 Technical information*. Retrieved from https://www.gl-assessment.co.uk/media/1343/cat4_extended_technical_information.pdf

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, *87*(2), 243–254. https://doi.org/10.1037//0021-9010.87.2.243

Hu, L., & Bentler, P. M. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

International Test Commission. (2001). ITC Guidelines on Test Use. *International Journal of Testing*, *1*(2), 93–114.

Jerrim, J. (2018). *How much does private tutoring matter for grammar school admissions? Education Datalab*. Retrieved from https://ffteducationdatalab.org.uk/2018/03/how-much-does-private-tutoring-matter-for-grammar-school-admissions/

Jesson, D. (2013). *The Creation, Development and Present State of Grammar Schools in England*. York. Retrieved from https://www.suttontrust.com/wp-content/uploads/2013/11/grammarsjesson.pdf

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal,

perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393–416. https://doi.org/10.1016/j.intell.2004.12.002

Kent County Council. (n.d.). Kent Test. Retrieved October 5, 2018, from https://www.kent.gov.uk/education-and-children/schools/school-places/kent-test

McDonald, A. S. (2001). The Prevalence and Effects of Test Anxiety in School Children. *Educational Psychology*, *21*(1), 89–101. https://doi.org/10.1080/01443410020019867

Schagen, I. A. N., & Schagen, S. (2003). Analysis of National Value-added Datasets to Assess the Impact of Selection on Pupil Performance. *British Educational Research Journal*, *29*(4), 561–582. https://doi.org/10.1080/0141192032000099379

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*(4), 363–374. https://doi.org/10.1016/j.intell.2005.11.004

Standards & Testing Agency. (2016). *How to convert key stage 2 raw scores to scaled scores*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/616977/2016_KS2_scaled_scores.pdf

Valerius, S., & Sparfeldt, J. R. (2014). Consistent g- as well as consistent verbal-, numerical- and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries. *Intelligence*, *44*, 120–133. https://doi.org/10.1016/j.intell.2014.04.003

Wiberg, M., & Sundström, A. (2009). A Comparison of Two Approaches to Correction of Restriction of Range in Correlation Analysis. *Practical Assessment, Research & Evaluation*, *14*(5). Retrieved from https://pareonline.net/getvn.asp?v=14&n=5

Table 1. Test score statistics by cohort (descriptive statistics based on available data are given first, and ML estimated statistics for the whole sample are given in parentheses)

	Cohort 1 (N_1 = 95)					Cohort 2 (N_2 = 99)				
Assessment	Min	Max	Med.	Mean	SD	Min	Max	Med.	Mean	SD
CAT4 (N=92)						(N=96)				
Verbal	67	136	104	102.80 (102.69)	15.32 (15.35)	64	141	108	104.10 (104.48)	17.10 (17.05)
Numerical	65	140	103	103.47 (103.47)	15.86 (15.98)	59	140	107	101.23 (101.51)	18.74 (18.59)
Non-verbal	69	138	105	104.95 (104.74)	15.12 (15.29)	69	141	110	106.58 (106.91)	18.25 (18.16)
Spatial	73	137	104	104.73 (104.68)	15.48 (15.60)	65	140	105	104.57 (104.97)	16.02 (16.03)
Kent Test (N=69)						(N=75)				
English	80	138	109	107.39 (103.32)	11.97 (13.47)	69	141	111	111.23 (105.08)	15.09 (18.25)
Mathematics	80	139	108	108.71 (103.29)	14.66 (16.83)	69	141	110	109.64 (101.15)	16.53 (21.90)
Reasoning	70	141	112	111.97 (106.22)	13.02 (15.73)	72	141	114	113.13 (106.71)	14.04 (17.74)
KS2 (N=95)						(N=99)				
Eng. Reading	83	118	104	103.14	8.06	83	120	108	104.79	9.77
Eng. Grammar	90	119	104	104.04	6.47	87	120	108	106.10	8.47
Mathematics	90	119	105	104.47	6.34	80	120	106	105.11	8.28

Table 2. Estimated correlations of assessment scores by cohort (results for Cohort 1 are below the diagonal, for Cohort 2 above the diagonal)

	CAT4				*Kent Test*			*KS2*		
Assessment	1	2	3	4	5	6	7	8	9	10
CAT4										
1 Verbal		.78	.78	.73	**.79**	.79	**.75**	**.79**	**.83**	.78
2 Quantitative	.68		.80	.77	.72	**.84**	.76	.76	.82	**.83**
3 Non-verbal	.74	.74		**.77**	.74	.80	**.77**	.69	.78	.79
4 Spatial	.70	.72	**.69**		.65	.71	**.78**	.67	.68	.72
Kent Test										
5 English	**.61**	.56	.48	.50		.82	**.79**	**.73**	**.79**	.75
6 Mathematics	.68	**.76**	.63	.61	.53		.84	.73	.84	**.87**
7 Reasoning	**.74**	.73	**.79**	**.64**	**.64**	.76		.69	.78	.78
KS2										
8 Eng. Reading	**.74**	.48	.61	.54	**.72**	.57	.68		**.81**	.76
9 Eng. Grammar	**.75**	.63	.63	.61	**.71**	.65	.75	**.75**		.82
10 Mathematics	.66	**.75**	.71	.69	.56	**.78**	.84	.67	.72	

Note: Correlations between tests of similar content are **bolded**; correlations between tests of similar format (knowledge-loaded or knowledge-free) are shaded.

Table 3. Goodness of fit for the alternative factor models by cohort

	General factor	*Fluid-crystallized*	*Verbal - numerical/figural*	*Verbal - numerical - figural*
degrees of freedom	35	34	34	32
Cohort 1				
χ^2 (*p*-value)	114.032 (<.001)	108.089 (<.001)	75.710 (<.001)	72.585 (<.001)
CFI	.888	.895	.941	.943
RMSEA	.154	.151	.114	.116
SRMR	.076	.076	.067	.058
BIC	6124.284	6122.895	**6090.516**	6096.498
Cohort 2				
χ^2 (*p*-value)	63.650 (.002)	60.636 (.003)	47.093 (.067)	41.133 (.129)
CFI	.969	.971	.986	.990
RMSEA	.091	.089	.062	.054
SRMR	.042	.044	.032	.030
BIC	6510.537	6512.118	**6498.575**	6501.805

Table 4. Standardized factor loadings for the alternative models (Cohort 1 / 2)

	General	*Fluid - crystallized*		*Verbal - numerical/figural*		*Verbal - numerical - figural*		
Assessment	g	fluid	cryst.	verbal	num./fig.	verb	numer.	figural
CAT4								
1 Verbal	.85/.88	.86/.88		.88/.90		.88/.90		
2 Quantitative	.83/.90	.83/.91			.85/.91		.86/.91	
3 Non-verbal	.83/.87	.85/.84			.84/.88			.84/.90
4 Spatial	.78/.81	.79/.92			.78/.82			.78/.84
Kent Test								
5 English	.68/.87		.71/.87	.78/.88		.76/.88		
6 Mathematics	.82/.92		.82/.93		.83/.93		.84/.94	
7 Reasoning*	.90/.88	.90/.91			.90/.89			.91/.90
KS2								
8 Eng. Reading	.77/.84		.79/.84	.84/.86		.84/.86		
9 Eng. Grammar	.83/.91		.85/.92	.88/.93		.88/.93		
10 Mathematics	.87/.90		.88/.91		.89/.91		.90/.92	

Note. * Kent Test reasoning component includes some verbal reasoning tasks; however, allowing this test to cross-load on the verbal factor in Verbal-numerical/figural and Verbal-numerical-figural models yielded insignificant loadings.

Table 5. Cross-tabulation of **combined** eligibility decisions (based on Kent Test and KS2) and admission decisions by cohort

		Admission decisions		
Cohort 1	*Eligibility*	H	G	
Kent Test	H	58	14	
	G	0	23	
	% correct	100%	62.2%	**85.3%**
KS2	H	51	11	
	G	7	26	
	% correct	87.9%	70.3%	**81.1%**
Cohort 2		*Admission decisions*		
	Eligibility	H	G	
Kent Test	H	46	14	
	G	0	39	
	% correct	100%	73.6%	**85.9%**
KS2	H	40	10	
	G	6	43	
	% correct	87.0%	81.1%	**83.8%**

Note. H= High school; G = Grammar school.

Table 6. Cross-tabulation of **aggregated** eligibility decisions (based on Kent Test and KS2) and admission decisions by cohort

		Admission decisions		
Cohort 1	*Eligibility*	H	G	
Kent Test	H	53	1	
	G	5	36	
	% correct	91.4%	97.3%	**93.7%**
KS2	H	44	7	
	G	14	30	
	% correct	75.9%	81.1%	**77.9%**
Cohort 2		*Admission decisions*		
	Eligibility	H	G	
Kent Test	H	45	1	
	G	1	52	
	% correct	97.8%	98.1%	**98.9%**
KS2	H	39	4	
	G	7	49	
	% correct	84.8%	92.5%	**88.9%**

Note. H= High school; G = Grammar school.

Table S1. Correlations of assessment scores by cohort based on available samples (results for Cohort 1 are below the diagonal, for Cohort 2 above the diagonal)

	CAT4				*Kent Test*			*KS2*		
Assessment	1	2	3	4	5	6	7	8	9	10
CAT4										
1 Verbal		.78[e]	.79[e]	.73[e]	.69[f]	.67[f]	.61[f]	.79[e]	.83[e]	.78[e]
2 Quantitative	.67[a]		.80[e]	.77[e]	.56[f]	.74[f]	.63[f]	.76[e]	.82[e]	.84[e]
3 Non-verbal	.74[a]	.73[a]		.77[e]	.61[f]	.69[f]	.66[f]	.69[e]	.78[e]	.79[e]
4 Spatial	.69[a]	.71[a]	.68[a]		.48[f]	.54[f]	.66[f]	.66[e]	.68[e]	.72[e]
Kent Test										
5 English	.50[b]	.41[b]	.31[b]	.37[b]		.71[g]	.68[g]	.59[g]	.67[g]	.58[g]
6 Mathematics	.57[b]	.69[b]	.53[b]	.51[b]	.40[c]		.73[g]	.54[g]	.71[g]	.76[g]
7 Reasoning	.62[b]	.63[b]	.70[b]	.56[b]	.51[c]	.67[c]		.52[g]	.63[g]	.63[g]
KS2										
8 Eng. Reading	.74[a]	.48[a]	.60[a]	.53[a]	.62[c]	.38[c]	.49[c]		.81[h]	.76[h]
9 Eng. Grammar	.75[a]	.62[a]	.63[a]	.60[a]	.61[c]	.51[c]	.63[c]	.75[d]		.82[h]
10 Mathematics	.65[a]	.74[a]	.70[a]	.68[a]	.43[c]	.70[c]	.78[c]	.66[d]	.72[d]	

Note: [a] Correlations are based on samples N=92; [b] N=66; [c] N=69; [d] N=95 for Cohort 1; and on samples [e] N=96; [f] N=72; [g] N=75; [h] N=99 for Cohort 2. All correlations are statistically significant, two-tailed, $p < 0.01$.