PeopleHawk

THE SCIENCE BEHIND PEOPLEHAWK:

The science behind PeopleHawk's **gamified cognitive assessments**

A focused approach to building scientifically-backed candidate profiles.

Gamified assessments

Gamified assessments, or game-based psychometric assessments, are an increasingly common recruitment tool to help recruiters screen numerous applications in search of the best candidates. These assessments usually involve a candidate playing a number of short rapid-response games and/or longer interactive scenarios, during which data points are gathered about that person. This provides businesses with a better measure of cognitive abilities than relying on a questionnaire or CV. Here we explore the three types of games that PeopleHawk has developed, and discuss their relevance to candidates and candidate selection.

Cognitive abilities

It is possible to combine scores on a number of different cognitive assessments or tests in order to estimate a person's level of general intelligence, which is closely related to inductive and deductive reasoning and to working memory (Cooper 2015). As discussed in other scientific papers general intelligence has also been shown to be a potent predictor of performance across a wide range of jobs and occupations (Thorndike 1985).

The general intelligence score, calculated by PeopleHawk is the average of the standardised scores candidates attain across all three of PeopleHawk's gamified assessments. This ensures that the three games are weighted equally, even though the standard deviation of their scores may differ.

The three PeopleHawk games provide measures of cognitive ability; that is to say they measure whether candidates can solve various puzzles, with a set time limit for each item. They also measure the level of performance at unfamiliar tasks which involve thinking and processing information.

Combined, cognitive ability tests and personality tests make up the majority of research on selection testing (Hough & Oswald, 2000;

Hough, Oswald, & Ployhart, 2001; Ones & Anderson, 2002; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Roth, Huffcutt, & Bobko, 2003;). It should be noted that PeopleHawk also provides a scientifically-validated personality quiz designed to assess candidate personality traits and work styles. Refer to PeopleHawk's scientific papers.

In a stand-alone basis Peoplehawk's three gamified assessments assess how well candidates perform various cognitive tasks and the relevance of each game to work performance is explained below.

Game 1: Anagrams

Anagrams involves rearranging letters to form a word and then finding a word related to it. For example, the letters APE can be rearranged to form the word PEA. Candidates are asked to solve the anagram, and then decide which of four answers is related to this solution. So, if the four possible answers were ARMCHAIR, SKY, BEAN and ELECTRIC, the correct answer would be BEAN, as beans are similar to peas.

Scores on a (short, 7-item) version of this game correlated 0.504 with the AH4 IQ test uncorrected for measurement errors when it was piloted as part of the 2003 “Test the nation” IQ test. This evidences that PeopleHawk’s Anagrams game measures a cognitive ability which is a component of general intelligence. In fact the game likely involves several cognitive skills. Familiarity with language is important – for example, if the letters N, I and g appear, the participant might suspect that the anagram ends with ING. Flexibility of thinking is also important, as it is necessary to discard strategies which do not work rather than sticking rigidly to them; “A GRIN” does not produce an anagram which ends in “ing”, for example.

All of PeopleHawk’s gamified assessments also involve speed of responding and this is known to be related to general intelligence; speed of problem solving is related to the number of problems answered correctly when people are given unlimited time to solve them.

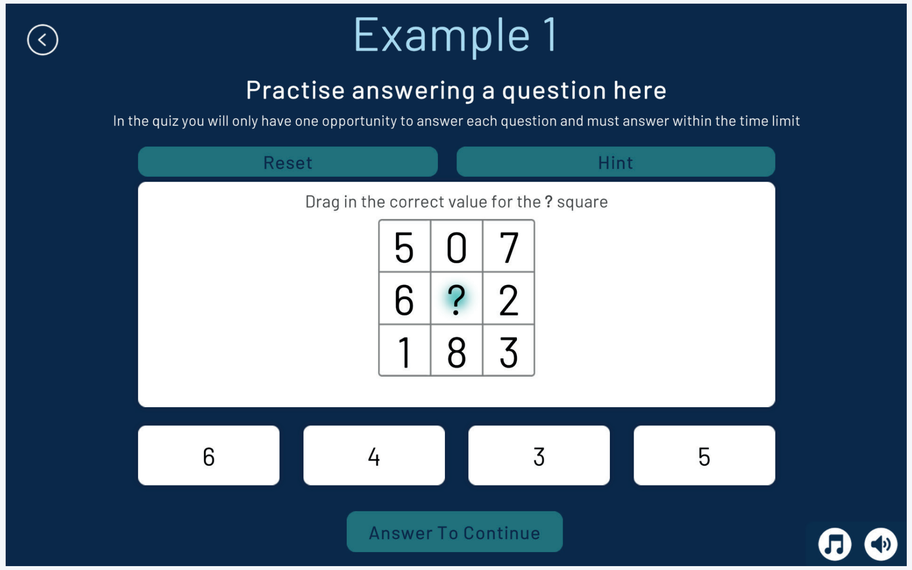
Game 2: Number Squares

Number Squares involves “magic squares” – 3 x 3, 4 x 4 or 5 x 5 squares holding numbers such that all the rows, columns and diagonals have the same total. The number in one of the cells is missing; participants choose which number needs to go in the missing cell from a list of four alternatives.

This is not particularly difficult and merely requires some mental arithmetic. However, the numbers in some of the other cells are obscured, so to solve the problem participants must work out which number should go in the obscured cell or cells and remember these whilst performing the addition and subtraction problems to determine the correct answer.

The ability to hold things in memory whilst performing other cognitive tasks is known as “working memory”. It has been extensively studied, and is known to be substantially related to general intelligence. This task also requires a considerable amount of planning (“executive functioning”) to work out how to solve the problem, as it is necessary to determine which of the obscured cells need to be solved as intermediate steps to producing the final answer.

A seven-item version of this task correlated 0.549 with the AH4 intelligence test when piloted for the 2006 “Test the Nation” IQ test, showing that it too is a reasonable measure of general intelligence.



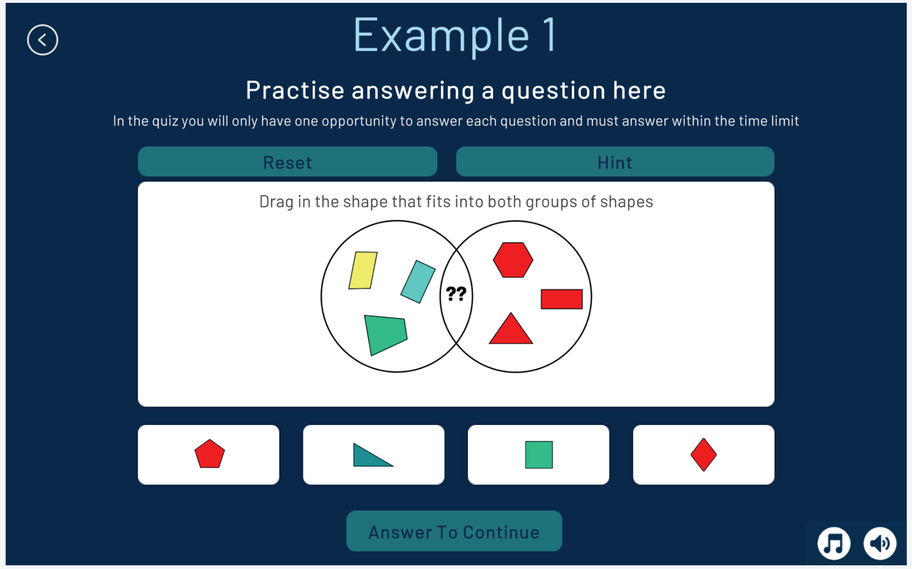
Game 3: Shapeshifters

Both of the previous games require some familiarity with words or arithmetic, but some intelligence tests investigate purely abstract reasoning. This approach has been incorporated into several commercial intelligence tests which focus on purely abstract, logical reasoning, using shapes rather than numbers or letters. This principle has been adopted in Shapeshifters.

The Shapeshifters game shows participants two sets of shapes. The first set of shapes will have something in common – they may all be red, for example. The second set of shapes will also have something in common – they may all be squares, for example. The participant's task is to identify which of four alternatives obeys both rules, and so belongs to both sets of shapes. Here the answer is a red square.

Finding the correct answer involves both inductive reasoning – inferring two rules from two sets of data; and deductive reasoning - deciding which answer obeys both rules. Both inductive and deductive reasoning involve logic and both are closely linked to general intelligence (Carroll 1993).

A seven-item version of this task correlated 0.661 with the AH4 intelligence test comparison detailed above.



What is the point of taking these games?

The candidate's perspective

PeopleHawk's three games are certainly enjoyable, but more importantly are informative measures of cognitive abilities which are known to be related to general intelligence. It is however important for candidates to complete the gamified assessments under optimal conditions i.e. treating them seriously and taking them when alert and in a quiet private place. Doing so will more likely better reveal the extent of a candidate's mental skills.

Rather than just giving a "score out of 15" for each game (which may be difficult or impossible to interpret), PeopleHawk compares individual candidate scores to those of other candidates who have also completed the same three gamified assessments. This means that candidates can see areas in which they comparatively excel and those which have comparatively some room for development.

However, it is important to realise that candidates are being compared to other applicants – most of whom will also be smart, ambitious graduates or professionals. So even if a candidate scores below the average (and half the people taking the test will do so!) remember that it is highly likely that this candidate will be well above the average score for the general population.

The employer's perspective

Cognitive assessments, such as Peoplehawk's games have been used for personnel guidance, screening and selection for over a century, with considerable success. Kanfer, Ackerman and Goff (1995) conclude that:

> *"Companies have unarguably saved billions of dollars by using ability tests to assure a merit-based selection process. . . . Overall, tests of intellectual abilities are the single most predictive element in employee selection . . . and certainly more valid than the use of personal interviews . . . in predicting training and on-the-job success. (p. 597)".*

Completing unfamiliar tasks such as PeopleHawk's gamified assessments reveals how well each candidate can formulate strategies to resolve novel problems and perform the necessary cognitive operations to find the correct solutions. For some tasks the planning element is fairly minor; for others, working out what needs to be done in order to solve the problem is what makes the task difficult. However, in both cases the unfamiliarity of the problems is of benefit as it allows one to separate potential from experience. In-basket exercises, which present applicants with job-specific problems and dilemmas to solve, necessarily confound experience with potential; a person can obtain a good score either because they have great experience in dealing with these types of problems but no great flair for dealing with the unfamiliar, or because they have little experience but superb problem-solving skills.

It transpires that assessments, such as Peoplehawk's games can be potent predictors of performance in many areas. Table 1 (taken from a meta-analysis by Hunter & Hunter, 1984) shows the correlation between test performance and job performance for large samples of workers. The larger the correlation, the better the level of prediction of performance. In short, a correlation of 0 implies that there is no link between test scores and job performance, whilst a correlation of 1 implies that the test can predict job performance with complete accuracy.

Table 1 shows that performance within each job category ***is*** related to test performance, and that these relationships are more substantial for senior staff than those in more routine roles. it is likely that this is because more senior roles tend to have a higher degree of complexity.

TABLE 1
Correlations between intelligence and job performance

Job category	Correlation between intelligence and job performance
Manager	0.53
Clerk	0.54
Salesperson	0.61
Protective professions worker	0.42
Trades and craft worker	0.46
Elementary industrial worker	0.37
Vehicle operator	0.28
Sales clerk	0.27

Source. Hunter and Hunter, 1984

The figures shown in Table 1 show the link between intelligence and job performance. Intelligence is estimated by simply averaging scores on a range of different tests, for it has been found (Thorndike 1985) that the reason why various tests measure job performance is because they all measure general intelligence to some extent. It matters less what the format of the tests are, but more importantly the fact that the tests are scientifically validated to measure general intelligence, leaning they are likely to predict performance.

Development of PeopleHawk's games

PeopleHawk's three gamified assessments were developed by its lead Scientist, Dr Colin Cooper. The three game formats came from the BBC "Test the Nation" IQ tests which were also developed by Dr Colin Cooper. However, the items themselves are unique to PeopleHawk and have deliberately been chosen to be more difficult than those used for the general population.

The advantage of this is that these formats were validated when building the Test the Nation tests by correlating shorter (7-item) and easier versions of the games with a commercial IQ test (the AH4 developed by Heim, Watts et al. 1970). This was administered under standard conditions using a representative sample of 300 adults from four regions of the UK. The correlations between those versions of the tests and the AH4 are shown in Table 2.

These substantial correlations show that the three games each measure cognitive abilities which are related to general intelligence.

TABLE 2

Correlations between easier versions of games and the AH4 test of general intelligence from three random samples of the UK population (N>300) tested during the development of the BBC "Test the Nation" IQ tests and DVD.

	Anagrams	Number squares	Shape shifter
Correlation with AH4	0.504	0.549	0.661

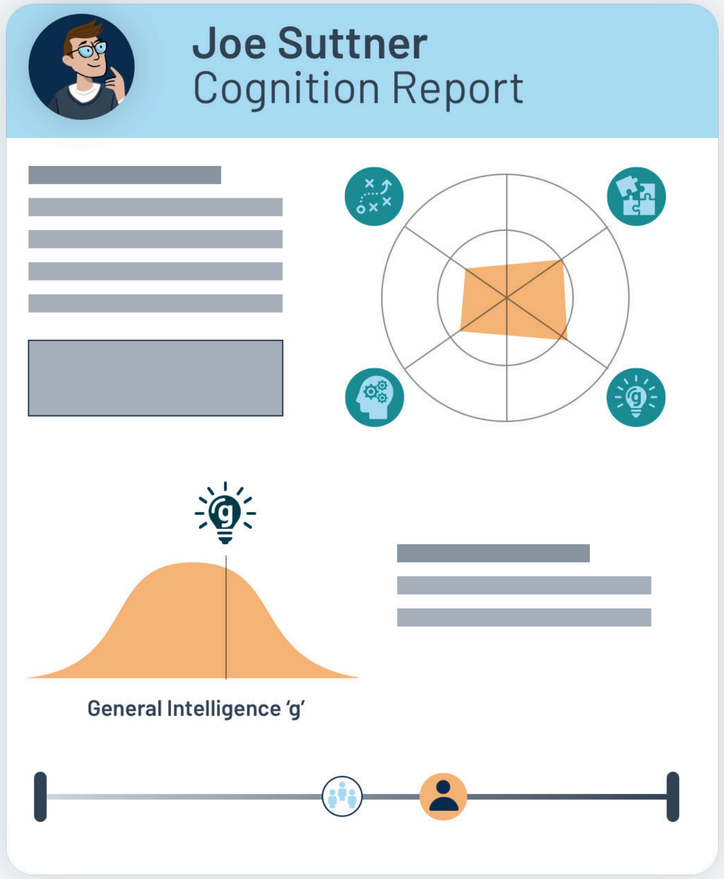
Further details of how the games were developed are given in Appendix A.

Summary Findings

Although the PeopleHawk games are user friendly and fun for candidates to complete, the scientific evidence underpinning them demonstrates that:

- They successfully measure cognitive abilities related to general intelligence;
- There is strong evidence that performance at games such as these predicts the level of performance in a wide range of jobs, particularly at senior levels; and
- The scores on each game show how each individual performs relative to others who have taken part.

As these participants are all likely to be high-achieving graduates, even a low score probably implies that candidates are above the average for the general population.



Appendix A: The detail behind PeopleHawk's game construction

Each game question/challenge provides candidates with four possible answers, only one of which is correct. Candidates are asked to guess if they are unsure of the correct answer, to ensure that anyone who fails to answer within the allocated time is not unfairly penalised (they have the same one-in-four chance of getting the item correct as if they had guessed at random).

Although the game formats are tried and tested as described above, new proprietary items were written specifically for the PeopleHawk games. These were administered to a sample of adults aged 21-35, living in the UK, USA or Ireland, and having a degree or higher qualification. Considerably more than 15 items have been written for each game. This is because it was necessary to produce many parallel versions to ensure that each participant was presented with a different version of the game. This was necessary to prevent candidates sharing the answers on social media, or remembering/writing the answers should they take a game on more than one occasion. We do not give details of the process or the number of items in the pool for each game as this is proprietary.

Factor analysis and item analysis was then performed on the entire pool of items to ensure that items were of appropriate difficulty, that all of the items in a particular game measured the same ability and that the amount of measurement error in a 15-item version of each game would be acceptable (reliability>0.7).

The correlations between the games were substantial (and statistically highly significant) as shown in Table 3.

TABLE 3
Correlations between long forms of the three games (N=194)

	Anagrams	Number squares	Shape shifter
Correlation with AH4	1.0		
NumberSquares	0.415	1.0	
ShapeShifters	0.418	0.379	1.0

As might be expected from Table 3, principal components analysis produced one component on which all three games had substantial loadings, as shown in Table 4. All the games have large loadings on what may be presumed to be a "general intelligence" factor.

TABLE 4

Results from a principal components analysis of the correlations between the three games (N=194).

	Loadings on 1st principal component
Anagrams	0.667
NumberSquares	0.668
ShapeShifters	0.649

After eliminating a few items (four or fewer per game) which were too easy, too hard, or showed low item-total correlations or factor loadings, extensive simulations were performed to ensure that candidates obtained near-identical scores no matter which 15-item version of the game they took.

To achieve this, 500 different versions of each 15-item game were generated and the scores of each of the sample study participants were calculated for each version of each game. The distribution of each person's scores on the 500 versions of a game was examined to ensure that scores were close to (+/- 1) the person's true score, as estimated from the long version of the game. 76% of the Number Squares variants, 74% of the ShapeShifters variants and 72% of the Anagrams met this criterion, showing that scores are highly similar no matter which version of the game a person is given. An example is shown below.

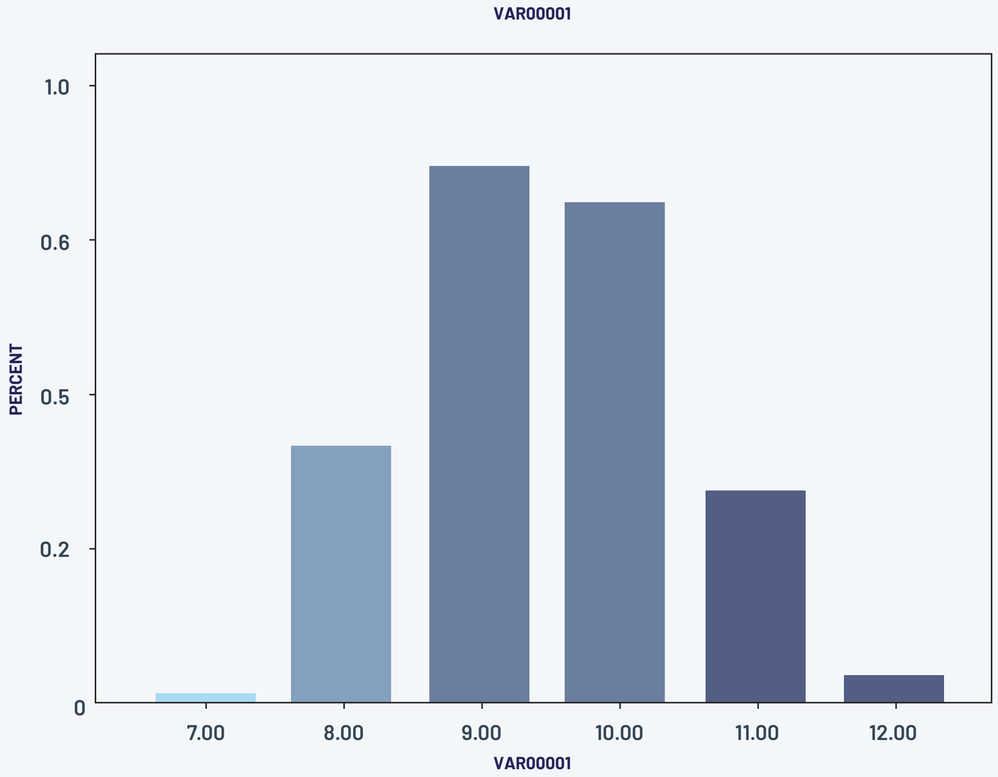


Figure 1. The scores of one randomly-selected person on 500 different versions of the ShapeShifters task.

Figure 1 shows the results of one of these simulations from one person who was chosen at random. 500 different short (15-item) versions of the ShapeShifters task were generated, and their scores (out of 15) on each version of the task were calculated. 84% of their scores were 9 +/- 1, indicating that they obtained a similar score no matter which version of the test they took.

The results were sometimes even more impressive. Figure 2 shows the results for another participant.

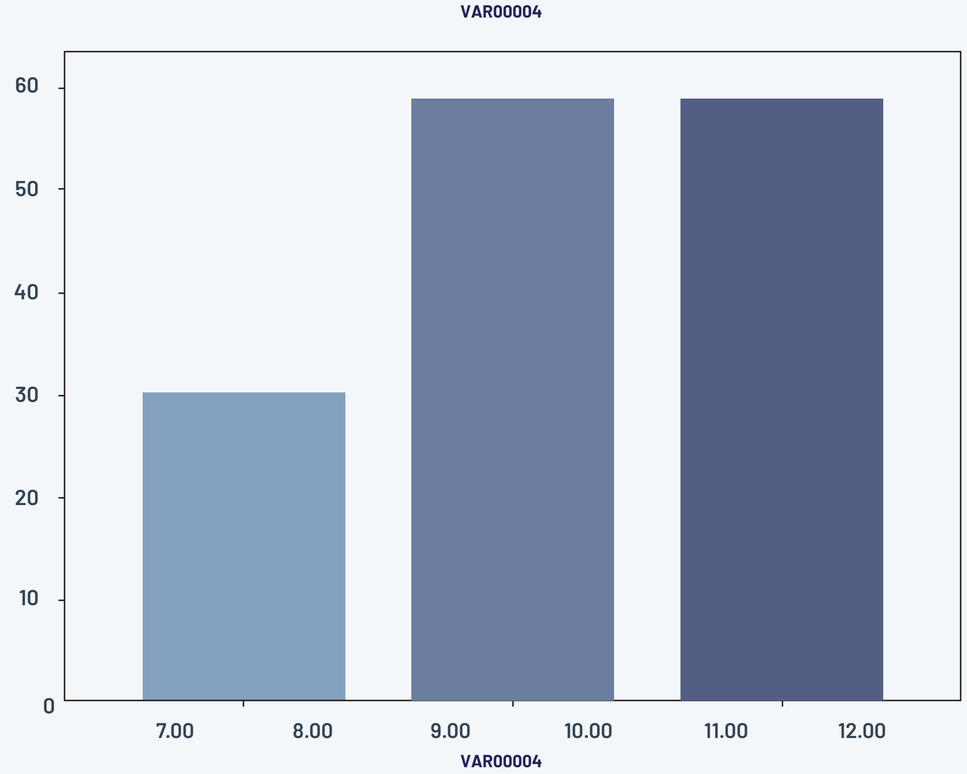


Figure 2. The scores of a second person on 500 different versions of the ShapeShifters task.

In Figure 2, all of this participant's scores on the 500 versions of the test are within 1 point of the median, 11.

Some researchers find gender differences when comparing the scores of adult males and females on tests of general ability (e.g., van der Linden, Dunkel et al. 2017; Irwing, 2012) although others find that these are trivial (Burgaleta, Head et al. 2012, Pezzuti and Orsini 2016). Whether or not a gender difference is found seems to depend on the types of problems that are administered. It is therefore important to check whether different genders perform at similar levels on the PeopleHawk games, to ensure that they do not show adverse impact when used as part of a selection procedure.

None of the three games showed adverse impart for gender; the means for males and females are shown in Table 5, below. Females slightly outperformed males on Anagrams, Shapeshifters and *g* (general intelligence) whilst males performed marginally better on Number Squares.

TABLE 5
Comparison of mean scores of males and females on random 15-item games, plus general intelligence

Descriptives

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Anagrams	Male	72	9.8611	3.23409	.38114	9.1011	10.6211	2.00	15.00
	Female	132	10.2424	2.79891	.24361	9.7605	10.7244	1.00	15.00
	Total	204	10.1078	2.95752	.20707	9.6996	10.5161	1.00	15.00
Numbers									
	Male	72	9.1667	3.50050	.41254	8.344	9.9892	1.00	15.00
	Female	132	9.0758	2.73686	.23821	.6045	9.5470	3.00	15.00
	Total	204	9.1078	3.02015	.21145	8.6909	9.5248	1.00	15.00
Shapes									
	Male	72	8.8472	3.55900	.41943	8.0109	9.6835	2.00	14.00
	Female	132	9.3561	3.22474	.28068	8.8008	9.9113	1.00	15.00
	Total	204	9.1765	3.34667	.23431	8.7145	9.6385	1.00	15.00
g									
	Male	72	-.0526	.90604	.10678	-.2655	.1603	-2.21	1.69
	Female	132	.0306	.67072	.05838	-.0849	.1461	-1.86	1.36
	Total	204	.0012	.76093	.05328	-.1038	.1063	-2.21	1.69

However the differences were tiny, and did not begin to approach statistical significance. This is shown in Table 6 (for a difference to be statistically significant, the number in the last column should be below 0.05).

TABLE 6
Descriptive statistics for random 15-item games, together with general intelligence

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Anagrams	Between Groups	6.774	1	6.774	.774	.380
	Within Groups	1768.854	202	8.757		
	Total	1775.627	203			
Numbers	Between Groups	.385	1	.385	.042	.838
	Within Groups	1851.242	202	9.165		
	Total	1851.627	203			
Shapes	Between Groups	12.062	1	12.062	1.077	.301
	Within Groups	2261.585	202	11.196		
	Total	2273.647	203			
g	Between Groups	.323	1	.323	.556	.457
	Within Groups	117.216	202	.580		
	Total	117.539	203			

REFERENCES

Burgaleta, M., Head, K., Alvarez-Linera, J., Martinez, K., Escorial, S., Haier, R. and Colom, R. (2012). "Sex differences in brain volume are related to specific skills, not to general intelligence." *Intelligence* 40(1): 60-68

Carroll, J. B. (1993). Human cognitive abilities: a survey of factor-analytic studies. Cambridge, Cambridge University Press.

Cooper, C. (2015). Intelligence and abilities : structure, origins and applications. Abingdon, Oxon ; New York, NY, Routledge.

Heim, A. W., Watts, K. P. and Simmonds, V. (1970). AH4, AH5 and AH6 Tests. Windsor, NFER.

Hunter, J. E. and Hunter, R. F. (1984). "Validity and utility of alternative predictors of job performance." Psychological Bulletin 96: 72-98.

Irwing, P. (2012). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences* 53: 126-131.

Kanfer, P. L., Ackerman, Y. M. and Goff, M. (1995). Personality and intelligence in industrial and organizational psychology. International Handbook of Personality and Intelligence. D. H. Saklofske and M. Zeidner. New York, Plenum.

van der Linden, D., Dunkel, C. S., and Madison, G. (2017). "Sex differences in brain size and general intelligence (g)." *Intelligence* 63: 78-88

Pezzuti, L. and Orsini, A. (2016). "Are the sex differences in the Wechsler Intelligence Scale for Children - Fourth Edition?" *Learning and Individual Differences* 45: 307-312

Thorndike, R. L. (1985). "The central role of general ability in prediction." Multivariate Behavioral Research 20: 241-254.