## Appendix A. Power calculations

We calculated 100 schools as the minimum necessary number to detect an effect of approximately 0.18 of a standard deviation in Key Stage 2 mathematics test scores. This calculation assumed:

(i) An intra-cluster correlation (ICC) of 0.15 at the school level[1]

(ii) Equal cluster sizes of 60 year 5 pupils per school[2]

(iii) 40 percent of the variation in KS2 maths test scores would be explained by baseline covariates[3]

(iv) 80 percent power for a 95 percent confidence interval

Table A1 provides estimates of the ICC for the actual sample of schools/pupils that took part in the study. Estimates are presented for baseline (KS1 average points score) and follow-up (KS2 maths) tests, when using either a fixed or random school level effect. The ICC for KS1 Average Point Scores (APS) was 0.08 when using a fixed effects model. The analogous ICC for KS2 maths was 0.13. In the results section, we illustrate that 45 percent of the variance in KS2 maths test scores can be explained by the baseline covariates. Using these figures in place of (i) and (iii) above, we calculate the minimum detectable effect in this trial was approximately 0.16.

**Table A1. Estimated inter-cluster correlation**

| | Fixed effect | Random effect |
|---|---|---|
| Key Stage 1 APS | 0.08 | 0.05 |
| Key Stage 2 Maths | 0.13 | 0.11 |

Note: Figures refer to the proportion of the variation in pupils' test scores occurring between school

[1] A value of 0.15 for the ICC was chosen after the team conducted an analysis of within and between school variation in key stage 2 test scores within the National Pupil Database.

[2] The figure of 60 pupils was based on the assumption of most recruited schools being two form entry, with each form containing 30 pupils.

[3] A value of 0.4 was chosen after the team conducted an analysis of the association between key stage 1 and key stage 2 test scores within the National Pupil Database.

## Appendix B. Estimated effect of the Chess in Schools intervention upon children's Key Stage 2 mathematics scores, by chess lesson 'quality'

Given differences in the size of our trial compared to others, one possible reason for differences in findings is the fact that it is harder to implement high-quality interventions to scale. We therefore investigate whether the impact of our trial varies by quality of the intervention. This information was captured within the qualitative process evaluation, with a basic 'tutor quality' measure created based upon children's reports of their enjoyment and engagement in the lessons. Children taught by 'low quality' tutors achieved Key Stage 2 scores slightly below the control group (-0.05 standard deviations) while children with 'medium quality' tutors scored a little higher than the control group (+0.11 standard deviations). However, there is no clear pattern of a 'dose-response' relationship, as the effect of having a high-quality tutor was essentially zero. Moreover, none of the estimates reach statistical significant at conventional levels. There is hence no evidence that the effect of the CSC intervention varied significantly by this particular measure of Chess lesson quality.

| Outcome | Effect size (95% CI) | p-value |
|---|---|---|
| **'low quality'** | -0.05 (-0.26 to +0.15) | 0.63 |
| **'medium quality'** | +0.11 (-0.07 to +0.29) | 0.25 |
| **'high quality'** | 0.00 (-0.27 to +0.26) | 0.99 |

Notes: Low, medium and high quality lessons based upon the proportion of pupils who enjoyed the chess tutors lessons. Effect sizes reported are relative to the reference group.

## Appendix C. Estimated effect of the Chess in Schools intervention upon children's Key Stage 2 mathematics scores, by measures of school quality

It may be the case that we did not find an impact of chess on attainment because our defined sample was particularly disadvantaged. We therefore investigate whether there was any differential impact of chess on attainment across school quality in our sample, using two different measures of school quality.

First, as noted in section 3, schools that participated in the trial were initially divided into ten separate strata based upon historical achievement data and the proportion of children eligible for Free School Meals. We have investigated how the estimated treatment effect varies across these strata, and whether there is any evidence of greater effects observed in higher-achieving, more affluent schools. We find little evidence that this is the case, with no consistent pattern of larger effect sizes within higher-achieving or less-deprived schools.

**Estimated effect of the Chess in Schools intervention upon children's Key Stage 2 mathematics scores, by randomisation strata**

| Strata | Number of pupils (schools) | Effect size | SE |
|---|---|---|---|
| Low achieving, high deprivation | 406 (12) | -0.21 | 0.13 |
| Low achieving, average deprivation | 487 (12) | 0.15 | 0.21 |
| Low achieving, low deprivation | 294 (7) | -0.32** | 0.12 |
| Average achieving, high deprivation | 305 (7) | -0.05 | 0.20 |
| Average achieving, average deprivation | 447 (11) | 0.15 | 0.21 |
| Average achieving, low deprivation | 561 (13) | 0.29* | 0.15 |
| High achieving, high deprivation | 346 (9) | -0.09 | 0.20 |
| High achieving, average deprivation | 537(10) | -0.17 | 0.14 |
| High achieving, low deprivation | 420 (11) | -0.10 | 0.10 |
| Late recruitment | 178 (7) | 0.00 | 0.14 |

Notes: * and ** indicate effect size statistically significant at the 10% and 5% levels respectively.

Second, in England, schools are regularly externally inspected and rated on a four-point scale (Outstanding, good, requires improvement and inadequate). These ratings are in part based upon inspectors' judgements of pupils behaviour, with previous research finding the impact of school-based interventions to vary by this factor (Jerrim and Vignoles 2016). It is thought that this likely to be due to the challenges of successfully implementing interventions within challenging schools.

**Estimated effect of the Chess in Schools intervention upon children's Key Stage 2 mathematics scores, by school inspection rating**

| Ofsted rating | Sample size pupils (schools) | Effect size | Standard error |
|---|---|---|---|
| **Overall grade** | | | |
| Outstanding | 493 (14) | -0.17 | 0.14 |
| Good | 2579 (65) | 0.04 | 0.09 |
| Requires improvement | 646 (17) | 0.01 | 0.13 |
| Missing data | 120 (3) | -0.59 | 0.32 |
| **Quality of teaching** | | | |
| Outstanding | 405 (12) | -0.06 | 0.15 |
| Good | 2667 (67) | 0.02 | 0.09 |
| Requires improvement | 646 (17) | 0.01 | 0.13 |
| Missing data | 120 (3) | -0.59 | 0.32 |
| **Behaviour of pupils** | | | |
| Outstanding | 1098 (30) | 0.03 | 0.11 |
| Good | 2395 (60) | 0.03 | 0.11 |
| Requires improvement | 225 (6) | 0.11 | 0.11 |
| Missing data | 120 (3) | -0.59 | 0.32 |

Notes: None of the estimates are statistically significant at the five per cent level.

Even in outstanding schools, with excellent teaching and well-behaved pupils, we still find no evidence that the CSC intervention had a positive impact upon pupil outcomes. Indeed, in contrast to Jerrim and Vignoles (2016), we find no evidence of heterogeneity in the effect by school inspection rating. We therefore believe that our focus upon lower-achieving schools is unlikely to be responsible for our failure to detect a positive treatment effect.

**Appendix D. Estimated effect of the Chess in Schools intervention upon children's Key Stage 2 scores, by subject dropped to make way for the chess lessons**

In our study, schools were allowed to choose how the hour of chess instruction would fit into their weekly timetable, though with the expectation this would be an art or humanities subject. Given this decision, 15 schools chose to drop an arts or humanities lessons, 13 used a mix of different (though not mathematics) lessons, nine were categorised as 'other' (including science, ICT, and physical education), seven dropped a mathematics lesson, while six did not receive the chess intervention (recall Figure 1).

Due to the small sample size within each category, most estimated treatment effects are statistically insignificant. However, the general direction of the point estimates suggests that schools which chose to drop an arts or humanities lessons tended to do slightly worse than the control group, while schools in the 'other' category tended to do slightly better. Moreover, there is no evidence that schools which replaced a mathematics lesson with a chess lesson did worse than the control groups.

| | Sample size pupils (schools) | Mathematics overall | | Mental arithmetic | |
|---|---|---|---|---|---|
| | | **Effect size** | **SE** | **Effect size** | **SE** |
| **Intervention Group (Ref: Control)** | 1926 (50) | | | | |
| Dropped mathematics | 333 (7) | 0.11 | 0.134 | 0.05 | 0.12 |
| Dropped arts/humanities | 683 (15) | -0.13 | 0.09 | -0.15* | 0.08 |
| Dropped a mix of subjects | 542 (13) | -0.01 | 0.09 | 0.08 | 0.08 |
| Dropped 'other' | 311 (9) | 0.26* | 0.14 | 0.19** | 0.09 |
| Crossed-over | 201 (6) | -0.18 | 0.12 | -0.19 | 0.12 |

| | Sample size pupils (schools) | Reading | | Science | |
|---|---|---|---|---|---|
| | | Effect size | SE | Effect size | SE |
| **Intervention Group (Ref: Control)** | 1926 (50) | | | | |
| Dropped mathematics | 333 (7) | 0.03 | 0.11 | -0.05 | 0.06 |
| Dropped arts/humanities | 683 (15) | -0.16** | 0.08 | -0.13 | 0.09 |
| Dropped a mix of subjects | 542 (13) | -0.02 | 0.09 | 0.05 | 0.07 |
| Dropped 'other' | 311 (9) | 0.11 | 0.15 | 0.05 | 0.10 |
| Crossed-over | 201 (6) | -0.29** | 0.08 | -0.07 | 0.07 |

Notes: Estimates refer to effect size (Cohen's D). Model 3 includes controls for baseline mathematics, reading, writing and science scores, gender and free school meal eligibility.