Elsevier Editorial System(tm) for Intelligence
Manuscript Draft

Manuscript Number: INTELL-D-14-00115R2

Title: Ageing trajectories of fluid intelligence in late life: the influence of age, practice and childhood IQ on Raven's Progressive Matrices

Article Type: Research Paper

Keywords: IQ, cognitive decline, ageing, practice, birth cohort, follow up study, Flynn effect, Matthew effect

Corresponding Author: Dr. Roger Staff, PhD

Corresponding Author's Institution: NHS-G

First Author: Roger Staff, PhD

Order of Authors: Roger Staff, PhD; Michael J Hogan, PhD; Lawrence J Whalley, FRCP(E) FRCPsych

Highlights

- We examined cognitive decline in two cohorts of healthy elderly people longitudinally born in 1921 and 1936
- We modelled decline and practice effects using a multilevel linear mixed modelling approach
- We estimate that on average IQ decreases annually by over one half point (age effect)
- Practice accounted for two IQ points on the second test occasion, no effect was observed thereafter
- Comparisons between birth cohorts suggest that the ‘Flynn’ effect influences our data
- We found no evidence of the ‘Mathew’ effect in late life

***Title Page with author details**

Ageing trajectories of fluid intelligence in late life: the influence of age, practice and childhood IQ on Raven's Progressive Matrices

Staff R.T.[1,2*], Hogan M.J.[3], Whalley L.J.[4]

Running title: The trajectory of fluid IQ in late life

[1] NHS –Grampian , Aberdeen United, Kingdom
[2] Aberdeen Biomedical Imaging Centre, University of Aberdeen, Aberdeen, United Kingdom.
[3] Department of Psychology, NUI, Galway, Ireland
[4] Institute of Applied Health Sciences, University of Aberdeen

[*]Corresponding Author
Aberdeen Royal Infirmary, NHS Grampian, Foresterhill, Aberdeen AB25 2ZD United Kingdom
TEL: UK-1224 552239
email: r.staff@abdn.ac.uk

Key words

IQ, cognitive decline, ageing, practice, birth cohort, follow up study, Flynn effect, Matthew effect

Abbreviations

RPM - Raven's Progressive Matrices
MHT – Moray House Test
P - Practice

Conflict of Interest

All authors have no conflicts of interest.

Funding

This work was supported by The Wellcome Trust, Alzheimer's Research UK and the Chief Scientist Office Scotland.

Ageing trajectories of fluid intelligence in late life: the influence of age, practice and childhood IQ on Raven's Progressive Matrices

Abstract

Background: Examining the trajectory of cognitive abilities in late life is difficult because of the long time required for significant change to occur and the confounding influences of baseline ability, cohort, drop out and practice.

Purpose: The aim is to describe cognitive trajectories in late life by estimating the influence of age and practice and by accounting for the potential confounding attributable to 'Flynn' and 'Mathew' effects.

Methods: We examine repeated measures of fluid intelligence (IQ) in 751 volunteers' ages 62 through 83 years sampled from the Aberdeen City 1921 and 1936 birth cohorts for whom archived childhood IQ data provided a rare insight into early life ability. Ageing trajectories in fluid intelligence were estimated using Raven's Standardised Progressive Matrices (RPM). Data were analysed using linear mixed models.

Results: We estimate that on average RPM decreases annually by over one half point (age effect). There is also an initial increase of about two points from the first to second test occasion the test is taken and this may be attributed to practice. Comparisons between birth cohorts suggest that the 'Flynn' effect influences our data. We found no evidence of the 'Mathew' effect in late life.

Conclusions: Cognitive trajectory of fluid ability in late life is a mixture of practice and decline. The influence of practice appears to be greatest after the first repeat testing. Modelling late life decline in this way will enable intervention studies to be performed in normal and prodromal dementia populations more efficiently.

1. Introduction

Examining the trajectory of cognitive abilities in late life is challenging due to the long time period required for significant change to occur and the confounding influence of baseline ability, age, cohort, drop out and practice. The objectives of this study are to quantify normative ageing trajectories and to identify potential predictors of these trajectories. In order to overcome challenges and obtain proper inferences, assumptions were made about the influence and mechanism of different predictors of cognitive ability change trajectories in late life.

One challenge in estimating longitudinal change in late life is that trajectories of change may be influenced by cohort effects. Cohort differences in socioeconomic circumstances, education, job complexity, leisure and cognitive pursuits may influence trajectories of change and result in different conclusions as regards normative patterns of change for different generations and historical periods. For example, the "*Flynn effect*" (Skirbekk, Stonawski, Bonsang, & Staudinger, 2013) describes a large-scale cohort effect that has been observed in most countries, whereby performance on IQ tests has improved with each decade of the 20th century. The '*Flynn effect'* has been estimated to be of the order of 3-4 IQ points for successive decades of the twentieth century (Brouwers, Van de Vijver, & Van Hemert, 2009).This observation that, coupled with other cohort differences, may influence patterns of age-related decline reported in cross-sectional and longitudinal studies . For example, using data from the Seattle Longitudinal Study Gerstorf and colleagues (Gerstorf, Ram, Hoppmann, Willis, & Schaie, 2011) reported less steep rates of cognitive aging between 50 and 80 years of age favoring the later born cohorts.

Estimating trajectories of cognitive change in late life is also challenging as a result of, dropout effects. Dropout occurs for the unavoidable reasons of death or illness and from voluntary withdrawal (P. Rabbitt, Lunn, & Wong, 2006). In addition, longitudinal studies by their nature involve repeated testing of individuals which introduces the additional influence of practice. That is, an individual may perform better because of familiarity with a test and testing environment. Practice effects of this type may be greater in individuals with higher original intelligence (IQ). Research using delayed growth models to evaluate patterns of skill acquisition suggests that baseline ability predicts rate of learning (Voelkle, Wittmann, & Ackerman, 2006). Longitudinal studies of IQ in children also provide evidence of the “Matthew effect” (MERTON, 1968; STANOVICH, 1986) whereby children with higher baseline IQ show more of an increase in IQ with repeated testing over time (SHAYWITZ et al., 1995). However, no study has examined if childhood IQ and/or adult ability predicts larger practice effects in the context of longitudinal analysis of cognitive change in older adults. In other words, it is unclear if the “Matthew effect” extends to old age trajectories of change.

Improvement with practice has been demonstrated by using a wide range of tests and in a variety of circumstances. Bartels et al (Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010)observed practice effects were stronger in the early phase than the late phase of high-frequency repetitive (1-6 monthly intervals) cognitive testing of healthy individuals. A plateau in performance followed with little or no effect of practice through repeating the test. Ferrer (Ferrer, Salthouse, Stewart, & Schwartz, 2004) sought to fit different shaped practice effects on longitudinal estimates of memory, space related abilities and speed of processing measures recorded at approximate yearly intervals. The shapes examined represent an initial practice model where the improvement is observed on the second testing occasion with negligible improvement thereafter. The second model represents a linear practice with equal amounts of improvements made on each testing occasion. Thirdly, the delayed practice model predicts small amounts of improvement are made initially and rates of improvement increase on each subsequent testing occasion. The form that this practice learning curve takes with our data and circumstances is unclear.

In addition to modelling the influence of practice, the separation of practice from the effect of age is a considerable analytical challenge. The need to model separate effects for age and practice are widely discussed (Ferrer et al., 2004; P. Rabbitt, Diggle, Holland, & McInnes, 2004; Salthouse, Schroeder, & Ferrer, 2004). When practice effects exist as a result of repeated assessments, ignoring the retest component will underestimate any age effect. When the frequency of testing (the interval between test and retest) is fixed, separating practice from decline is difficult since the unit of time and the interval of practice are essentially equivalent. McArdle and Woodcock

(McArdle & Woodcock, 1997) have proposed a model capable of handling different test intervals that permits the separation of practice and age.

Trajectories of cognitive change in late life can differ for different cognitive ability measures. Fluid ability is the capacity to think logically and solve problems in novel situations, independent of acquired knowledge (Horn & Cattell, 1966). Fluid ability typically peaks in young adulthood and may steadily decline thereafter.

Here, we use linear mixed modelling to estimate the influences of childhood IQ, age and retest practice on fluid ability in a well characterized sample of older adults with archived IQ scores at age 11 years and whose fluid ability was tested between ages 62 and 83 years on up to five occasions. We use a previously published age/practice model (Ferrer, Salthouse, McArdle, Stewart, & Schwartz, 2005) and hypothesize that (1) age has an overall significant negative effect on ability and (2) practice a significant positive influence on performance on Raven's Standardised Progressive Matrices (RPM), a robust test of fluid intelligence. We include IQ test scores at age 11 years to extend the Ferrer et al model and hypothesise (3) that higher childhood intelligence and/or greater adult premorbid ability predicts greater practice improvement. This hypothesis essentially tests the presences of the '*Matthew effect*' in late life. By including childhood ability, we adjust for potentially confounding influence of the "*Flynn effect"* on our data. In addition we also examine the influences of different practice models on our results. Consistent with previous research we further hypothesise that (4) ability at entry into the study is associated with greater practice improvements, and (5) older age at entry into the study is associated with less practice improvement. Overall, the approach adopted in this study will allow us to examine the relationship between childhood ability, ability at entry into the study, age at entry into the study, decline with age and practice improvement.

2. Methods

2.1. The Sample

All data were provided by the Aberdeen Birth Cohorts of 1921 and 1936, an extended description of samples recruitment and data acquisition is given in Whalley et al (Whalley et al., 2011). Following guidelines provided by the local ethics of research committee (University of Aberdeen and NHS-Grampian), volunteers gave written informed consent to a longitudinal observational study of brain ageing and health. Briefly, RPM data were collected longitudinally from two identically recruited cohorts between January 1998 and December 2011. For each cohort the number of recruitment and follow up measurements for each study year is shown in **Table 1**. RPM scores were recorded up to five times in each participant between ages 62 and 83 years. All data were collected between January 1998 and Aug 2013. Total data for combined cohorts comprised 751 individuals for whom 2153 RPM scores were recorded. The interval between consecutive tests ranged from a few months to greater than 10 years, with a median of 24 months. This wide range of interval was not planned and was a result of the varied amount of time the participants were able to devote to the project at each test point and the resources available for collection. Each individual was screened for dementia at recruitment and before testing using Mini Mental State examination (MMSE>25). The total sample contained 361 (48.1%) females.

2.2. Measures

Childhood intelligence data were kindly provided by the Scottish Council for Research in Education (SCRE) from data archived in the Scottish Mental Surveys of 1932 or 1947. All children born in either 1921 or 1936 and at school in Scotland on 1st June 1932 or 4th June 1947 when aged about 11 years sat a group-administered intelligence test (The Moray House Test (MHT)) (I. J. Deary, Whalley, & Crawford, 2004). Current adult fluid intelligence was assessed using Raven's Standardised Progressive Matrices (RPM) (Raven, 2000a), in which each test item consists of a matrix of geometric patterns with one missing cell, and the task for the participant is to select the best completion of the missing cell from a set of alternatives. This is a 60-item, timed (maximum 20 minutes allowed) fluid intelligence test of non-verbal abstract reasoning. The choice of RPM was based on its psychometric properties and reliance on non-verbal abstract reasoning. The RPM test manual describes studies from a wide age range and on different cultural groups from clinical and normal populations, and attest to the reliability (Alpha coefficient = 0.91) and validity (Raven, 2000a) of the measure.

2.3. Statistics

Before linear mixed modelling, the raw RPM scores and the MHT scores were standardised to a mean of 100 and a standard deviation of 15 on an IQ-like scale. Age at testing was measured as the number of years after the 60^{th} birthday. Practice was modelled by occasion of testing so that on first testing, occasion 1 has no influence of practice.

In order to examine the influence of age and practice (*P*) only on RPM, we used a linear mixed model as used by Ferrer et al 2004 (model 1a) and summarised in Equation 1. In this model, the beta values for the intercept, age and P were all random effects. That is to say $\beta_{0,j}$, $\beta_{1,j}$ and $\beta_{2,j}$ were all allowed to vary randomly between subjects (j).

$$RPM_{i,j} = \beta_{0,j} + \beta_{1,j} Age + \beta_{2,j} P + e_{i,j} \qquad 1$$

i= The occasion recorded

j= The subject

Age= The age-60 of subject when RPM_i was recorded on occasion *i*

P=The hypothesised practice effect,(e.g linear assumption P=*i*-1 for *i*=1, 2, 3, 4, 5)

$e_{i,j}$=The residual.

We extended this model to include childhood IQ at age 11 years as measured by the MHT (model 1b) and summarized in equation 2. This additional variable was also treated as a random effect.

$$RPM_{i,j} = \beta_{0,j} + \beta_{1,j} Age + \beta_{2,j} P + \beta_{3,j} MHT + e_{i,j}$$

In each case the model was fitted using the MLwin software from the University of Bristol (Rasbash, Steele, Browne, & Goldstein, 2012)]. Age was entered into the model as the number of years beyond participants' sixtieth birthday. This allowed the intercept (time or age = 0) to be interpreted as an estimate of adult pre-morbid ability. In all models we assume that practice effects can be estimated as linear over time and number of tests. That is to say, we assume that an individual improves by the same amount each time the test is taken. In order to examine this assumption in relation to improvement due to practice we repeated model 1a but replaced the linear practice assumption with an 'initial' model of practice (model 2a) where the participant improves after the first occasion and has negligible improvement thereafter (P=0 when i=1, P=1 when i≠1). We then extended this model to include childhood intelligence (model 2b). Similarly, we then replaced the practice assumption with a 'delayed' model (model 3a) where there is little improvement at first followed by delayed improvement after more than one occasions (P=0, 0.5, 1, 2, 4 when i=1,2,3,4,5 respectively). We then extended this model to include childhood intelligence (model 3b).

TABLE 1 ABOUT HERE

3. Results

The age and summary (raw) RPM and MHT scores for participants are shown in Table 1. The raw mean values for the RPM measures suggest improvement with age and improvement with occasion. The tables also indicate that individuals tested on more occasions had better RPM and MHT scores at baseline, indicating that the more cognitively able remained in the study. The table also shows that the 1936 cohort had a superior ability when compared to the 1921 cohort on both at entry into the study and in childhood, providing support for a “Flynn effect” difference between these cohorts. A ‘spaghetti plot’ for each individual split by the number of samples they provided is shown in Figure 1.

FIGURE 1 ABOUT HERE

TABLE 2 ABOUT HERE

The findings from all models are in Table 2 and show that age has a significant negative fixed effect on old age IQ and practice has a significant positive fixed effect on IQ, thus supporting hypothesis 1 and 2 and as expected childhood ability has a significant positive effect. The variance estimates for each model produces a varying picture. On all occasions the Age and Practice variances were significantly different from zero. The variance associated with childhood ability (MHT) did not differ from zero for all models. The intercept variance was significant for models that did not contain the MHT score (the ‘a’ models). Only the linear practice model (1b) demonstrated a variance significantly different from zero for the intercept.

Table 2 also shows a significant negative covariance between β_1 (age) and β_2 (practice) for all models This is in the same direction as those found by Ferrer et al for processing speed, and are consistent with hypothesis (5), which stated older age at entry into the study is associated with less practice improvement . With the exception of model 1 without the childhood ability no covariance was found between the intercept β_0 and β_1 (age).   Similarly no significant covariance was found between the intercept β_0 and β_2 (practice) and therefore does not support the hypothesis (4) that ability at entry into the study is associated with greater practice improvements. All of the covariances involving MHT were non- significant. In particular, absence of a significant covariance between β_2 (practice) and β_3 (MHT) provides no support for our hypothesis (3) that higher childhood intelligence predicts greater practice improvement and, therefore, provides no evidence of a “Matthew effect”.

The b models showed that the childhood IQ variance was not significantly different from zero and there were no significant co-variances between it and the other coefficients indicating that its influence is fixed. Therefore, while childhood IQ predicted ability in older adulthood it was found to have no significant association with the trajectory with age or practice.

Examining the *-2*loglikelihood* values for each model the best fit was found using the initial practice model 2. This would indicate that improvement due to practice is seen after the first repeat test and little is seen after subsequent testing.

A potential weakness of our study is the incomplete data acquisition and therefore some of the inferences may be based in part on extrapolation rather than actual data. In order to support the robustness of our findings we repeated model 2, the initial improvement model using data from participants who provided at least 3 RPM measures (Model 4). The results are shown in table 2. The fixed effects remained unchanged with minor differences between the β values. There is some change in the patterns of significant in the random effects. In particular the *intercept-Age c*ovariance was no longer significant in model ‘a’ and the *Age-P* co-variance was no longer significant in model ‘b’.

4. **Discussion**

As hypothesized, age and practice had significant influences on late life RPM test scores, such that with increasing age RPM scores declined each year by about half of an IQ point but increased when the test is repeated at the second occasion. Our results indicate that this improvement is found after the first retest and that little improvement is seen due to practice if the test is administered subsequently. The directions of these effects are the same as those previously reported (Ferrer et al., 2004) who modelled memory and processing speed as outcome measures. The inclusion of childhood IQ in the model did not significantly alter these effects with β_1 *and* β_2 not differing significantly between the ‘a’ and ‘b’ versions of each practice model. In addition the results show that childhood IQ did not co-vary with either the age or practice effect (β_1, β_2). That is to say, childhood IQ did not affect the rate of decline associated with aging or improvement with practice. Similarly, childhood IQ did not co-vary with the intercept and its variance was not significantly different from zero. This would imply that the influence of childhood IQ is a fixed effect on fluid intelligence in late life and that the variance between individuals is negligible. Taken together, these results support the hypothesis that childhood IQ influences late fluid intelligence (I. Deary, Whalley, Lemmon, Crawford, & Starr, 2000). However, childhood IQ has no direct influence on trajectory.

In this study, we assumed three different practice models and found that the initial practice model to be a superior fit to our data. These results differ from those published by Ferrer et al examining the trajectory of memory and processing speed (Ferrer et al., 2004) and who found superior fits for delayed and linear practice models. There are, however, considerable differences between our study, the Ferrer study and other longitudinal studies examining practice effects which examine effects derived from repetitions closer in time than presented here and using test and populations significantly different from ours. In reality, the appropriate practice model is probably test, interval and sample dependent and although the initial model may not be appropriate for all subjects, in these data it provides a reasonable approximation.

This study extends previous work on part of this data that found a significant relation between lower childhood MHT and greater decline in cognitive ability in later life (Bourne, Fox, Deary, & Whalley, 2007). The data and analytical model explored here are superior to our earlier analysis because more data are now available and by our inclusion of practice effects in the analytical model. These findings do not completely agree with our earlier conclusions.

Although we find that age is kinder to the initially more able (I. Deary, MacLennan, & Starr, 1998); our observations suggest that greater childhood IQ is of value only when, as in most cases, it leads to in superior adult cognitive ability. The absence of significant co-variance between MHT and the effects of age is consistent with this interpretation. That is, there appears to be no latent protective influence, represented by childhood IQ, not explained by adult pre-morbid ability. This would imply that converting early cognitive potential into adult ability protects an individual. This conversion may be facilitated by factors such as education or cognitively stimulating employment. These are of course commonly used proxies of cognitive reserve (Staff, Murray, Deary, & Whalley, 2004). The mechanism or mechanisms of hypothesized *"cognitive reserve"* are unclear (Staff, 2012) but may be a result of increased neural plasticity in individuals with more education and more occupational complexity (Karp et al., 2009) and consequently increased ability to compensate for neuropathological changes in the ageing brain. Similarly, it may well be brought about by individual differences in personality traits such as openness (Hogan 2012) where individuals with a high level of this trait are more active and engaged resulting in the maintenance of ability, a 'use it or lose it' hypothesis. An alternative proposed mechanism suggests that use-dependent processes afford cognitive protection. That is the development of more efficient (or at least different) processing mechanisms make individual less vulnerable to the effects of ageing (Waiter et al., 2008).

The “*Flynn effect*” predicts the cohort born in 1936 would have superior childhood IQ at age 11 and in late life. The MHT and RPM raw values in Table 1 indicate that to be the case although these values and different ages are probably distorted by survival at age 11 and both practice and survival effects in late life. Survival bias in the older (1921) sample may be brought about by those with higher childhood IQ surviving to an age when they were recruited (Waiter et al., 2008) and were, therefore, more often included in our study; this would suppress the Flynn effect gains. If this were the case, then we would expect the differences between those born in 1921 and 1936 to be smaller at age 11 than that predicted by the ‘*Flynn effect*’ which predict around 3 IQ points per decade or around 4.5 points between our 15 year interval (Raven, 2000b). Standardizing the MHT scores indicated a difference between the cohorts of 3.7 points. This is slightly smaller than expected and may be brought about by survival and selection bias discussed above. Late life comparisons indicate a significantly greater difference between the cohorts, comparing the cohorts at age 77; where there is overlap in data we find a difference of 10.4 raw RPM points or 16.5 IQ points, which is surprisingly large. The survival effect, describe above would suggest that the 1921 participants were of greater ability than those who did not take part in our study. The practice effect would suggest that the 1936 scores include the improvement due to repeat testing since all of them had experienced the test at least once before age 77. This possible confounder would probably result in increasing the differences between the cohorts and therefore 16.5 IQ point at age 77 is an overestimate. This is significantly greater than expected from the ‘*Flynn effect’* and our potential selection bias, although Flynn himself

(Flynn, 2012) has observed massive gains of between 5 and 25 points in a single generation. The marked contrast the age 11 Flynn effect gains with the age 77 effect gains demonstrated for the first time, within the same individuals that cohort gains may vary over the life span.

Analysing our cohorts separately (without childhood IQ) we found the age slope in the 1921 cohort to be (β_1 -0.551 SE 0.224) compared with (β_1 -0.356 SE 0.073) in the 1936 cohort. Similarly we found the practice slope to be (β_2 1.925 SE 0.858) in the 1921 cohort and (β_2 2.163 SE 0.571) in the 1936 cohort. These results would indicate that there is no significant difference between the two trajectories. However; this comparison is of limited value because the cohorts are not matched for age. There may be a cohort effect on trajectory, but this may be obscured by the age differences.

These results provide no support for the '*Matthew effect*' acting in late life, that is we found no significant covariance between the β_2 (Practice) and β_3 (MHT) or β_0 (the intercept). The '*Matthew effect*' has predominantly been observed in the cognitive domain in early life and has been clearly demonstrated in the acquisition of reading skills (STANOVICH, 1986). Little has been written about this effect in late life. It may well be that this effect is only realized in early life when the brain is more plastic and the advantages of better abilities are exploited.

One of the limitations of our study concerns the level of drop out leading to incomplete data acquisition. 42.7 % of the data points were not collected and only 15.7% of subjects provided all of the required data. Table 1 indicated that the more cognitively able were more likely to return. This pattern of dropout is well-known from other work (HULTSCH, HERTZOG, SMALL, MCDONALDMISZCZAK, & DIXON, 1992). Similarly, Rabbitt et al (P. Rabbitt et al., 2006) showed that individuals who completed a 20-year longitudinal study demonstrated little or no change and that those who dropped out showed significant cognitive decline. The existence of such selective dropout in our sample suggests that follow-up scores would be inflated when compared to a complete data set, resulting in less evidence for decline than would be seen in the population. This over-representation of the more able in the follow-up measures is compounded further by their over representation in the whole sample as the more cognitively able were more likely to volunteer (I. J. Deary, Whiteman, Starr, Whalley, & Fox, 2004). Attrition due to drop out is unavoidable; particularly in elderly samples were death accounts for a significant proportion of incomplete data. Similarly, illness may also account for a proportion of drop out (Matthews et al., 2004). In addition, those with some insight into their declining cognition may also choose not to continue, because of a reluctance to expose or confirm their decline. Collectively these factors are likely to produce an underestimate of decline. This may also inflate the practice effect among individuals whose impression is that they ‘did better’ on occasion 2 and are more likely to return. An additional limitation is that all the models examined here assume that incomplete data were missing at random. While this it is a rarely met assumption in longitudinal studies and probably not the case here, the assumption is found to perform well in most practical situations (Collins, Schafer, & Kam, 2001). The sub analysis of

Model 4 supports this assumption in terms of the fixed effects. However, the more subtle random effects, such as significant variances in the slope for age and practice and the covariances previously observed were not supported. The reduction in the estimates of variance in Model 4 would be expected given the selection criteria. Specifically, our results suggest that those declining faster are more likely to dropout and would not provide data across 3 occasions. The reduction in variance combined with the reduced sample size and the bias introduced by the selection process would limit the power to detect the more subtle variable relationships in our data. The use of Model 4's 'complete case' data filtering method has the additional drawback that it assumes that the data is missing completely at random, that is, there is no relationship between the missing data and the variables in the model (Rubin 1976). This is clearly not the case, with those who perform better as children more likely to be retained in the study, a pattern that is also observed when examining the RPM score at entry and retention in the study across occasions. Models 1-3 make no such assumption and therefore probably provide a more reliable and valid estimate of population effects (Graham 2009).

We were unable to record exactly reasons for not returning to the study. This was not our key focus which was to examine the influence of practice and childhood IQ on decline with age. Reasons for not returning in similar studies have been examined (Der, Allerhand, Starr, Hofer, & Deary, 2010). They found that progression to dementia, death, and refused testing due to illness are all factors that influenced cognitive trajectory in late life using a linear mixed model without accounting for practice. Although these factors are of interest they are unlikely to affect our core findings.

The variation in sampling interval across all participants and within participants has allowed us to discriminate between practice and age effects. If we applied a vigorous fixed interval each test occasion would be equivalent to a unit of time (age) which would make differentiation between the two influences more difficult. Measuring and accounting for practice effects in the context of decline in dementia and old age has design implications for clinical trials. The use of fixed time points, for example baseline, 3 months, 6 months etc, in the design, which is common place in drug studies does not provide sufficient variance between the practice occasion and time variables. A strictly acquired data set will have testing on occasion 2 always at 3 months and occasion 3 at 6 months and so on. This would make it difficult to separate the two influences. A more subtle design may well vary (perhaps randomised) the interval between repeat measures in order to model and separate such effects.

The key strength of our study is the almost unique availability of childhood IQ data, and late life fluid intelligence measure. Together, these are rarely found in a single sample and the finding that childhood IQ does not provide a latent influence on late life trajectory, independent of late life pre-morbid ability is relevant to trial design.

Disappointing results from recent drug trials in symptomatic dementia have switched research focus to interventions starting during the dementia prodrome. Well-timed, early interventions in the course of illness now appear more likely to succeed (Morris, 2005) with strong arguments for clinical trials in pre-symptomatic individuals at greatest dementia risk (Sperling et al., 2011). In the absence of overt dementia, however, trials are hindered by smaller than expected rate cognitive decline over time in those with little or no impairment which are probably confounded by practice. Understanding, the relationship between 'real' decline and the influence of practice will improve the sensitivity of such studies by; sample enrichment, identifying those who are likely to decline.

5. Highlights

- We examined cognitive decline in two cohorts of healthy elderly people longitudinally born in 1921 and 1936
- We modelled decline and practice effects using a multilevel linear mixed modelling approach
- We estimate that on average IQ decreases annually by over one half point (age effect)
- Practice accounted for two IQ points on the second test occasion, no effect was observed thereafter
- Comparisons between birth cohorts suggest that the 'Flynn' effect influences our data
- We found no evidence of the 'Mathew' effect in late life

		Occasion *i*					Total 21/**36**	Total	Mean RPM (Stnd Dev)	
		1	2	3	4	5			1921	1936
Age (years)	62	**1**					**1**	1		23.0 (-)
	63	**56**					**56**	56		36.8 (9.3)
	64	**161**					**161**	161		34.9 (9.0)
	65	**192**	**17**				**209**	209		35.7 (8.3)
	66	**67**	**141**				**208**	208		36.5 (8.4)
	67	**13**	**131**	**11**			**155**	155		36.9 (9.1)
	68	**5**	**65**	**121**			**191**	191		37.6 (8.1)
	69	**1**	**13**	**133**			**147**	147		37.4 (8.4)
	70		**1**	**21**			**22**	22		38.4 (7.7)
	71		**1**				**1**	1		27.0 (-)
	72				**8**		**8**	8		39.6 (7.7)
	73			**2**	**47**		**49**	49		35.9 (7.5)
	74		**3**	**5**	**69**		**77**	77		39.5 (6.4)
	75	1	**4**	**18**	**45**		1/**67**	68	38.0 (-)	40.3 (9.0)
	76	30					30	30	31.9 (9.8)	
	77	157	10/**2**	**7**	**17**	**61**	167/**87**	254	26.9 (8.8)	37.3 (9.9)
	78	31	70	**1**	**1**	**5**	101/**7**	108	28.0 (8.8)	41.9 (4.4)
	79	21	46	20			87	87	29.0 (8.7)	
	80	13	27	65	12		117	117	29.5 (8.6)	
	81	2	15	38	56	1	112	112	29.5 (8.4)	
	82			10	20	44	74	74	30.7 (8.3)	
	83			4	7	7	18	18	28.4 (8.7)	
Total 21/**36**		255/**496**	168/**378**	137/**319**	95/**187**	52/**66**	707/**1446**			
Total		751	546	456	282	118	2153			
Mean	MHT 1921 (Stnd Dev)	38.7 (12.7)	39.9 (12.2)	40.2 (11.9)	41.6 (44.6)	44.5 (10.1)				
	MHT 1936 (Stnd Dev)	42.1 (13.4)	43.5 (12.8)	44.5 (12.1)	46.3 (11.1)	48.7 (10.9)				
	All (Stnd Dev)	41.0 (13.3)	42.4 (12.7)	43.2 (12.1)	44.7 (11.3)	46.8 (10.6)				
Mean	RPM 1921 (Stnd Dev)	27.5 (9.0)	29.0 (8.8)	29.8 (8.3)	29.9 (8.3)	30.4 (8.2)				
	RPM 1936 (Stnd Dev)	35.4 (8.9)	37.3 (8.5)	37.7 (8.4)	38.7 (8.1)	38.0 (8.2)				
	All (Stnd Dev)	32.7 (9.6)	34.7 (9.4)	35.2 (9.1)	35.7 (9.1)	34.6 (9.0)				

Table 1

The number of the participants acquired at each occasion and age. The numbers in bold are from the 1936 cohort

		Model 1a	Model 1b	Model 2a	Model 2b	Model 3a	Model 3b	Model 4a	Model 4b
		Fixed effects							
Weights	β_0 *intercept*	104.6 (0.8)	52.6 (3.5)	102.7 (0.8)	51.5 (3.5)	104.1 (0.8)	51.6 (3.5)	106.8 (0.9)	59.4 (4.81)
	β_1 *Age*	-0.72 (0.08)	-0.67 (0.07)	-0.50 (0.06)	-0.47 (0.06)	-.60 (0.08)	-0.53 (0.08)	-0.64 (0.07)	-0.59 (0.08)
	β_2 *P*	1.62 (0.23)	1.38 (0.23)	2.48 (0.37)	2.25 (.39)	1.87 (0.33)	1.42 (0.33)	2.41 (0.43)	2.20 (0.42)
	β_3 *MHT*		0.51 (0.03)		0.50 (0.03)		0.51 (0.03)		0.45 (0.04)
		Random effects							
Variance	σ^2_0 *intercept*	212.3 (25.9)	737.9 (338.6)	187.8 (18.7)	517.7 (328.6) n.s.	215.1 (33.2)	626.0 (345.5) n.s	143.3 (23.3)	877.2 (502.4)n.s
	σ^2_1 *Age*	0.83 (0.30)	1.07 (0.26)	0.55 (0.15)	0.63 (0.14)	0.80 (0.36)	1.02 (0.33)	0.11 (0.23)n.s.	0.21 (0.22) n.s.
	σ^2_2 *P*	4.76 (1.93)	5.89 (1.77)	10.7 (6.3)	12.0 (6.3)	8.40 (3.2)	8.35 (2.90)	7.18 (6.80) n.s.	7.86 (6.61) n.s.
	σ^2_3 *MHT*		0.04 (0.03) n.s		0.03 (0.03) n.s.		0.03 (0.03) n.s.		0.05 (0.04) n.s.
Co-variance	$\rho_{0,1}$ *intercept-Age*	-7.42 (3.11)	-10.4 (6.8) n.s.	-3.81 (1.56)	-3.42 (5.63) n.s.	-7.25 (4.02)	-8.37 (7.55) n.s.	-0.27 (2.36) n.s.	-4.11 (7.66) n.s.
	$\rho_{0,2}$ *intercept-P*	13.2 (8.83) n.s.	26.9 (23.2) n.s	4.61 (8.01) n.s.	16.4 (36.9) n.s.	20.0 (16.6)	30.9 (37.4) n.s.	-8.01 (9.91) n.s.	-20.1 (42.5) n.s.
	$\rho_{0,3}$ *intercept-MHT*		-5.0 (3.2) n.s.		-3.27 (3.13) n.s.		-3.96 (3.22) n.s.		6.36 (4.58) n.s.
	$\rho_{1,2}$ *Age-P*	-1.43 (0.73)	-2.02 (0.64)	-1.16 (0.74) n.s.	-1.58 (0.72)	-1.94 (1.16)	-2.47 (1.06)	0.33 (.94) n.s.	-0.144 (0.91) n.s.
	$\rho_{1,3}$ *Age-MHT*		0.02 (0.06) n.s		0.00 (0.05) n.s.		0.00 (0.06) n.s.		0.03 (0.06) n.s.
	$\rho_{2,3}$ *P-MHT*		-0.08 (0.20) n.s.		-0.08 (0.34) n.s.		-0.07 (0.31) n.s.		0.12 (0.38) n.s.
Residual	$e_{i,j}$	34.2 (1.6)	34.0 (1.5)	33.7 (2.0)	33.0 (2.0)	35.0 (1.6)	34.4 (1.6)	31.1 (2.2)	30.6 (2.2)
Fit	*-2*loglikelihood*	15877	15645	15227	14541	15242	14553	9917	9807

Table 2

The values in the brackets are standard error. All values were significantly different from zero unless specified (n.s.). Model 1 - the linear practice model, Model 2 - the initial practice model, Model 3 - the delayed practice model, Model 4 – the initial practice model with participants that provided the first three complete occasions. *i*= The occasion recorded; *j*= The subject; *Age*= The age of subject when RPM_i was recorded on occasion *i*; *P*=The the hypothesised practice unit; *MHT*= The Moray House Test score age 11 years; $e_{i,j}$=The residual

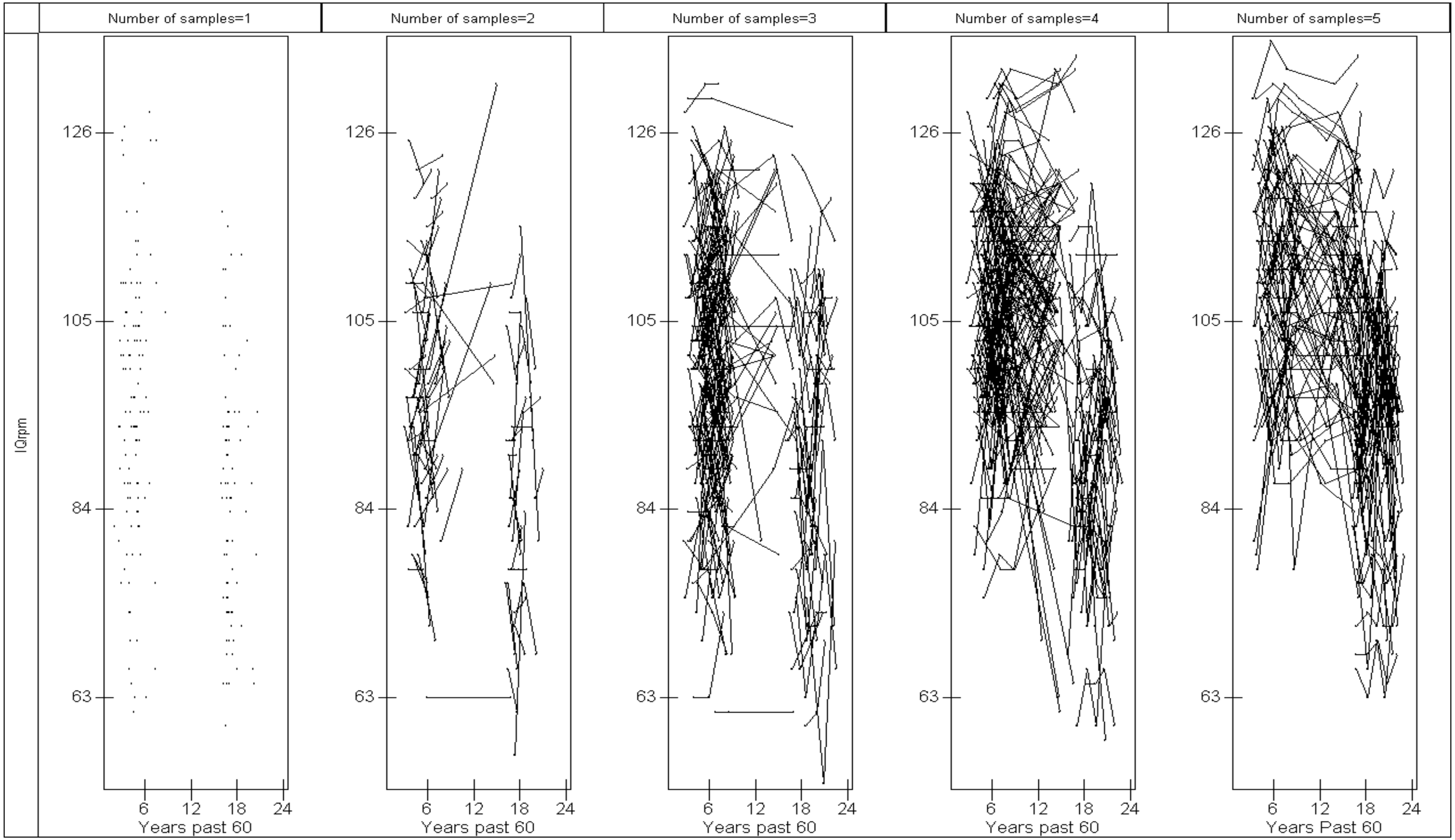

Figure 1

A spaghetti plot of the IQrpm scores for each participant split by the number of samples they provided

References

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *Bmc Neuroscience, 11*, 118. doi:10.1186/1471-2202-11-118

Bourne, V. J., Fox, H. C., Deary, I. J., & Whalley, L. J. (2007). Does childhood intelligence predict variation in cognitive change in later life? *Personality and Individual Differences, 42*(8), 1551-1559. doi:10.1016/j.paid.2006.10.030

Brouwers, S. A., Van de Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in raven's progressive matrices scores across time and place. *Learning and Individual Differences, 19*(3), 330-338. doi:10.1016/j.lindif.2008.10.006

Collins, L., Schafer, J., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi:10.1037//1082-989X.6.4.330

Deary, I. J., Whalley, L. J., & Crawford, J. R. (2004). An 'instantaneous' estimate of a lifetime's cognitive change. *Intelligence, 32*(2), 113-119. doi:10.1016/j.intell.2003.06.001

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*(1), 130-147. doi:10.1037/0022-3514.86.1.130

Deary, I., MacLennan, W., & Starr, J. (1998). Is age kinder to the initially more able?: Differential ageing of a verbal ability in the healthy old people in edinburgh study. *Intelligence, 26*(4), 357-375. doi:10.1016/S0160-2896(99)00005-7

Deary, I., Whalley, L., Lemmon, H., Crawford, J., & Starr, J. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 scottish mental survey. *Intelligence, 28*(1), 49-55. doi:10.1016/S0160-2896(99)00031-8

Der, G., Allerhand, M., Starr, J. M., Hofer, S. M., & Deary, I. J. (2010). Age-related changes in memory and fluid reasoning in a sample of healthy old people. *Aging Neuropsychology and Cognition, 17*(1), 55-70. doi:10.1080/13825580903009071

Ferrer, E., Salthouse, T. A., McArdle, J. J., Stewart, W. F., & Schwartz, B. S. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychology and Aging, 20*(3), 412-422. doi:10.1037/0882-7974.20.3.412

Ferrer, E., Salthouse, T., Stewart, W., & Schwartz, B. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging, 19*(2), 243-259. doi:10.1037/0882-7974.19.2.243

FLYNN, J. (2012). Are we getting smarter? Rising IQ in the twenty-first century. Cambridge University Press. ISBN 978-1-107-60917-4.

Gerstorf, D., Ram, N., Hoppmann, C., Willis, S. L., & Schaie, K. W. (2011). Cohort differences in cognitive aging and terminal decline in the seattle longitudinal study. *Developmental Psychology, 47*(4), 1026-1041. doi:10.1037/a0023426

Graham J. W. Missing Data Analysis: Making It Work in the Real World. Annu. Rev. Psychol. 2009. 60:549–576. doi: 10.1146/annurev.psych.58.110405.085530

Hogan, M.J., Staff, R.T. Bunting, B.P. Deary, I.J. Whalley, L.J. Openness to experience and activity engagement facilitate the maintenance of verbal ability in older adults. Psychol Aging. 2012 Dec; 27(4):849-54. doi: 10.1037/a0029066.

Horn, J. O. H. N. L., & Cattell, R. A. Y. M. O. N. D. B. (1966). Age differences in primary mental ability factors. *J Gerontol, 21*((2)), 210-220.

HULTSCH, D., HERTZOG, C., SMALL, B., MCDONALDMISZCZAK, L., & DIXON, R. (1992). Short-term longitudinal change in cognitive performance in later life. *Psychology and Aging, 7*(4), 571-584. doi:10.1037//0882-7974.7.4.571

Karp, A., Andel, R., Parker, M. G., Wang, H., Winblad, B., & Fratiglioni, L. (2009). Mentally stimulating activities at work during midlife and dementia risk after age 75: Follow-up study from the kungsholmen project. *American Journal of Geriatric Psychiatry, 17*(3), 227-236.

Matthews, F., Chatfield, M., Freeman, C., McCracken, C., Brayne, C., & MRC CFAS. (2004). Attrition and bias in the MRC cognitive function and ageing study: An epidemiological investigation. *Bmc Public Health, 4*, 12. doi:10.1186/1471-2458-4-12

McArdle, J., & Woodcock, R. (1997). Expanding test-retest designs to include developmental time-lag components. *Psychological Methods, 2*(4), 403-435. doi:10.1037/1082-989X.2.4.403

MERTON, R. (1968). Matthew effect in science. *Science, 159*(3810), 56-&. doi:10.1126/science.159.3810.56

Morris, J. (2005). Early-stage and preclinical alzheimer disease. *Alzheimer Disease & Associated Disorders, 19*(3), 163-165.

Rabbitt, P., Diggle, P., Holland, F., & McInnes, L. (2004). Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences, 59*(2), P84-P97.

Rabbitt, P., Lunn, M., & Wong, D. (2006). Understanding terminal decline in cognition and risk of death - methodological and theoretical implications of practice and dropout effects. *European Psychologist, 11*(3), 164-171. doi:10.1027/1016-9040.11.3.164

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2012). *A User's guide to MLwiN, v2.26.* Centre for Multilevel Modelling, University of Bristol.

Raven, J. (2000a). The raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48. doi:10.1006/cogp.1999.0735

Raven, J. (2000b). The raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48. doi:10.1006/cogp.1999.0735

Rubin DB. 1976. Inference and missing data. *Biometrika,* 63:581–92. doi:10.1093/biomet/63.3.581

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology, 40*(5), 813-822. doi:10.1037/0012-1649.40.5.813

SHAYWITZ, B., HOLFORD, T., HOLAHAN, J., FLETCHER, J., STUEBING, K., FRANCIS, D., & SHAYWITZ, S. (1995). A matthew effect for iq but not for reading - results from a longitudinal-study. *Reading Research Quarterly, 30*(4), 894-906. doi:10.2307/748203

Skirbekk, V., Stonawski, M., Bonsang, E., & Staudinger, U. M. (2013). The flynn effect and population aging. *Intelligence, 41*(3), 169-177. doi:10.1016/j.intell.2013.02.001

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., . . . Phelps, C. H. (2011). Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimers & Dementia, 7*(3), 280-292. doi:10.1016/j.jalz.2011.03.003

Staff, R. T. (2012). Reserve, brain changes, and decline. *Neuroimaging Clinics of North America, 22*(1), 99-+. doi:10.1016/j.nic.2011.11.006

Staff, R., Murray, A., Deary, I., & Whalley, L. (2004). What provides cerebral reserve? RID C-6297-2009 RID E-9019-2011. *Brain, 127*, 1191-1199. doi:10.1093/brain/awh144

STANOVICH, K. (1986). Matthew effects in reading - some consequences of individual-differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360-407. doi:10.1598/RRQ.21.4.1

Voelkle, M. C., Wittmann, W. W., & Ackerman, P. L. (2006). Abilities and skill acquisition: A delayed growth curve approach. *Learning and Individual Differences, 16*(4), 303-319. doi:10.1016/j.lindif.2006.01.001

Waiter, G. D., Fox, H. C., Murray, A. D., Starr, J. M., Staff, R. T., Bourne, V. J., . . . Deary, I. J. (2008). Is retaining the youthful functional anatomy underlying speed of information processing a signature of successful cognitive ageing? an event-related fMRI study of inspection time performance RID C-8951-2011 RID E-9019-2011 RID C-6297-2009. *Neuroimage, 41*(2), 581-595. doi:10.1016/j.neuroimage.2008.02.045

Whalley, L. J., Murray, A. D., Staff, R. T., Starr, J. M., Deary, I. J., Fox, H. C., . . . Crawford, J. R. (2011). How the 1932 and 1947 mental surveys of aberdeen schoolchildren provide a framework to explore the childhood origins of late onset disease and disability. *Maturitas, 69*(4), 365-372. doi:10.1016/j.maturitas.2011.05.010

***Response to Reviewers**

Reviewer #1: Flynn: Excellent rewrite- accept

Thank you for your assessment and your comments throughout the review process.

Reviewer #3: Reviewer #3
This revised manuscript has been modified in response to the prior reviews, and is now clearer in many respects.

Thank you for your assessment and your comments throughout the review process.

I still have two concerns about issues I raised in the original review.

One has to do with the comparability of results from the sub-sample with complete data across 3 occasions. The authors state that there was "some change in the patterns of significant (sic) in the random effects," but Table 2 reveals that there was no longer significant variance in either the age or practice effects, or in their covariances with the intercept. In my view these findings lead to questions about the robustness of the results in the other models, and I feel that at minimum they warrant further discussion.

We previously included the following text in response to Reviewer #3 comment#2 on our original submission;

"The sub analysis of Model 4 supports this assumption in terms of the fixed effects. However, the more subtle random effect was not supported by the sub analysis and their robustness is uncertain. It may well be the case that this violation may have limited the power to estimate random effects childhood IQ after adjustment for age and practice".

We have replaced this text with following text in response to the comment of this occasion;

"The sub analysis of Model 4 supports this assumption in terms of the fixed effects. However, the more subtle random effects, such as significant variances in the slope for age and practice and the covariances previously observed were not supported. The reduction in the estimates of variance in Model 4 would be expected given the selection criteria. Specifically, our results suggest that those declining faster are more likely to dropout and would not provide data across 3 occasions. The reduction in variance combined with the reduced sample size and the bias introduced by the selection process would limit the power to detect the more subtle variable relationships in our data. The use of Model 4's 'complete case' data filtering method has the additional drawback that it assumes that the data is missing completely at random, that is, there is no relationship between the missing data and the variables in the model [Rubin 1976]. This is clearly not the case, with those who perform better as children more likely to be retained in the study, a pattern that is also observed when examining the RPM score at entry and retention in the study across occasions. Models 1-3 make no such assumption and therefore probably provide a more reliable and valid estimate of population effects [Graham 2009]."

The second issue concerns the explanation for not including additional cognitive measures in this manuscript. In their resubmission letter the authors state that "examination of other cognitive domains is an obvious extension to our work and one which we hope to explore in the near future." I suspect that the manuscript would make more of a unique contribution if additional measures were to be included because not only could relations be examined with more measures, but possible coupling of change in different cognitive measures could be explored. As noted by one of the other reviewers, the contribution of the present manuscript could be considered modest, but this would likely not be the case if more of the existing data were to be included in this report.

We are grateful for the suggestion proposed here, but examining the, “possible coupling of change in different cognitive measures”, would require a separate specific hypothesis about the coupling of the change process in this context. It is our view that the addition of other measures is beyond the scope of the original intended focus, which was to examine the intellectual trajectory in our defined sample and to examine the influence of early life ability on that trajectory. We would like to thank all of the reviewers for their comments as they have substantially improved our manuscript, however, the inclusion of additional cognitive measure as suggested by Reviewer#3 will result in a manuscript that is inconsistent with our original intentions. If the reviewer is concerned about serial publication with the same data set to optimize our publication count, this is not the practice of our group or any individuals in it. The core business of this group currently is to identify the reasons for individual differences in late life decline rather than exploring the more theoretical concepts such as domain coupling. The reviewer describes our manuscript as a ‘modest’ contribution, which may be a reasonable assessment; however, the impact of our work using this unique data set can only be evaluated once it is in the public domain. ”