Intelligence xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Intelligence

Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups

Timothy C. Bates [a,b,*], Shivani Gupta [a]

[a] Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK
[b] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form 15 November 2016
Accepted 15 November 2016
Available online xxxx

Keywords:
Collective intelligence
Group IQ
IQ
Gender
Communication
Group psychology
Administrative behavior

ABSTRACT

What allows groups to behave intelligently? One suggestion is that groups exhibit a collective intelligence accounted for by number of women in the group, turn-taking and emotional empathizing, with group-IQ being only weakly-linked to individual IQ (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Here we report tests of this model across three studies with 312 people. Contrary to prediction, individual IQ accounted for around 80% of group-IQ differences. Hypotheses that group-IQ increases with number of women in the group and with turn-taking were not supported. Reading the mind in the eyes (RME) performance was associated with individual IQ, and, in one study, with group-IQ factor scores. However, a well-fitting structural model combining data from studies 2 and 3 indicated that RME exerted no influence on the group-IQ latent factor (instead having a modest impact on a single group test). The experiments instead showed that higher individual IQ enhances group performance such that individual IQ determined 100% of latent group-IQ. Implications for future work on group-based achievement are examined.

1. Introduction

While humans form groups and value group membership (Haidt, 2007; Lewis & Bates, 2010), this has typically been understood in terms of obedience and loyalty adaptations maximizing goal completion (Simon, 1997). Recently, Woolley, Chabris, Pentland, Hashmi, & Malone (2010) reported a new possible benefit of group work: the emergence of a collective intelligence factor largely unrelated to individual IQ. They reported that people working on complex problems in groups show a strong general-ability or IQ factor, with significant differences between groups on this factor. Surprisingly, group-IQ, or "collective intelligence" (C) as they termed it was "*not strongly correlated with the average or maximum individual intelligence of group members but is correlated with the average social sensitivity of group members, the equality in distribution of conversational turn-taking, and the proportion of females in the group.*" Woolley et al. (2010, p. 686). These findings were subsequently argued to warrant a "*seismic shift in how we study groups*" (Woolley & Malone, 2011, p. 2).

As the editors of Nature (Nature Editorial, 2016) recently commented regarding replication studies "*researchers must make more of them, funders must encourage them and journals must publish them.*" (p. 373). Here, in three independent samples, we therefore tested these hypotheses, and contrasted these against the hypothesis that group-IQ predominantly reflects individual cognitive ability.

For some time, it has been known that work-groups whose teammembers have higher IQ out-perform teams of less-able members (Devine & Philips, 2001). Against this background, Woolley et al. (2010) asked whether groups themselves exhibit a general-factor of intelligence, if this might be distinct from individual IQ, and, if so, what the origins of such a collective intelligence might be. Woolley et al. (2010) assessed individual IQ using either Raven's matrices (Raven, Raven, & Court, 1998) or the Wonderlic Personnel Test – a brief multiple-choice measure of intelligence (Wonderlic & Hovland, 1939). Social sensitivity was assessed using the Reading the Mind in the Eyes (RME) task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). RME involves subjects viewing images of expressive faces, masked to show only the eye region, and choosing which of four words plotted around the image best describes the depicted emotion. To assess group-IQ, subjects were allocated to small groups and performed tasks including brainstorming, matrix reasoning, moral reasoning, planning a shopping trip, and collaborative text editing (see Woolley et al., 2010 Supplementary Tables S1a and S3b for range of tasks used in their study 1 and study 2). These reflect the McGrath (1984) task circumplex – an established taxonomy for measuring group performance. The four quadrants of the circumplex are: (1) 'Generate' – development of new ideas; (2) 'Choose' – tasks that require definitive correct answers; (3) 'Negotiate' – resolving conflicts of interest or points of view; and (4) 'Execute' – performance and psychomotor tasks. A confirmatory factor analysis

* Corresponding author at: Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK.
E-mail address: tim.bates@ed.ac.uk (T.C. Bates).

http://dx.doi.org/10.1016/j.intell.2016.11.004
0160-2896/

(CFA) indicated that a single latent factor accounted for 31–35% of test variance. Surprisingly, individual IQ accounted for just 3% of group-IQ variance.

Turning to the causes of this group-IQ factor, Woolley et al. (2010) reported a significant ($r = 0.23$, $p = 0.007$) correlation with percent females in the group. Variance in turn-taking during communication similarly correlated positively and significantly with group-IQ. In multiple regression models, these factors were displaced by social sensitivity (RME), which was the best predictor of group-IQ. Their conclusion was that a distinct form of collective intelligence exists which can solve complex problems independent of the IQs of individual group members. If social sensitivity enables a collective problem solving ability not limited by conventional cognitive ability of the group members this would clearly be of profound importance, especially given that simply increasing female participation and encouraging turn-taking might allow us to increase collective ability.

Given the ubiquitous importance of group activities (Simon, 1997) these results have wide implications. Rather than hiring individuals with high cognitive skill who command higher salaries (Ritchie & Bates, 2013), organizations might select-for or teach social sensitivity thus raising collective intelligence, or even operate a female gender bias with the expectation of substantial performance gains. While the study has over 700 citations and was widely reported to the public (Woolley, Malone, & Chabris, 2015), to our knowledge only one replication has been reported (Engel, Woolley, Jing, Chabris, & Malone, 2014). This study used online (rather than in-person) tasks and did not include individual IQ. We therefore conducted three replication studies, reported below.

2. Study 1

Based on Woolley et al. (2010), we set out to confirm replication of the following hypotheses. First, in a battery of group-tasks, a single factor should account for a substantial portion of variance in scores. Second, individual IQ would be a poor (path-coefficient ≤ 0.20) predictor of group-IQ. Third, number of women in the group would predict group-IQ. And fourth, social sensitivity would strongly account for variance in group-IQ, explaining for the predicted apparent association of number of women with group-IQ, and greatly exceeding any effect of individual cognitive ability.

2.1. Method

2.1.1. Participants

Seventy-two (41 females, 31 males) student participants were recruited using Facebook and university class e-mail lists. The age range of participants was eighteen to twenty-four years of age. One subject was in full time employment. Subjects were offered a £50 prize for the best performing group. For collective IQ testing, these 72 subjects were formed into 26 groups (as described below).

2.1.2. Materials

Individual IQ was assessed using the Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998), a standardized test of general fluid reasoning capacity. Participants were given 10 min to complete as many as possible of the odd-numbered items of set-II of this test and were scored for total correct, as in Woolley et al. (2010) study 1.

Note: When conducting these studies, we expected a group-IQ factor to emerge independent of IQ, and wished to consider alternative models for collective cooperation based on personality and moral psychology. For this reason, subjects in study one completed the NEO-FFI Five Factor Personality Inventory (Costa & McCrea, 1992) while in studies 2 and 3, subjects completed the Moral Foundations Questionnaire (Graham, Haidt, & Nosek, 2009) and a measure of psychopathy (Christie & Geis, 1970). We had no need to explain results based on these ancillary scales, and for this reason they are not analyzed or presented here.

Individual Social Sensitivity was assessed using the "Reading the Mind in the Eyes" test (Baron-Cohen et al., 2001). This 35-item test involves viewing pictures of emotional expressions cropped to just the eye-region, and picking the correct descriptor from among three foil words.

2.1.3. Group-IQ assessment tasks

Tasks used to assess group-IQ were 1) Brainstorming, 2) Group Ravens, 3) Plan Shopping Trip, and 4) Architectural Design. These were selected based on their factor loadings in Woolley et al. (2010) Study 1.

Brainstorming draws on Quadrant 1 of the McGrath circumplex, and involved each group writing down as many possible uses for both a brick and of a paperclip, with 5 min given for each item. Responses were scored following Wilson, Guilford, and Christensen (1953) and based on the number of uses generated, originality and the frequency in comparison to other groups.

Group Ravens (Quadrant 2) involved groups completing as many of the even numbered questions in set II of Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998) as they could in 10 min, scored for the number of correctly answered items. In *group planning* (Quadrant 2) each group planned a shopping trip as if they were members of a household buying groceries. Members each had a different list of items they needed to collect while sharing a single car. Various constraints were put in place, for example certain items like milk would spoil after 45 min. There are better and worst places for members to buy different items for example some shops sell better quality items and some shops sell cheaper items. The goal was to make a plan in which they purchased as many high-value items as possible while adhering to the constraints. Teams received 5 min of instructions and then had 15 min to complete the task.

The architecture task (Quadrant 3: negotiation) followed Woolley et al. (2010) and involved each group designing and building a model house from a limited set of building blocks. Essential features were one door, two windows and a roof. Teams were given 5 min to organize, and 15 min to complete the task. Structures were scored on size, durability and aesthetic quality, and received penalties if they failed to conform to the essential criteria. We thank a reviewer who asked us to note that in Woolley et al. (2010) teams built a house, garage, and pool, receiving 15 min of instructions, 10 min of planning, and 20 min to build; here, teams built a house with one door, two windows and a roof with 5 min to plan and 15 min to build.

Finally, a computerized game of checkers was administered. Used as a criterion task by (Woolley et al., 2010), we selected this task based on its factor performance, and analyzed it (equivalently) as additional manifest measure of collective ability. Group members played checkers against a computerized opponent. Members were first familiarized with the rules of the game, then given time for a short practice match and lastly played one test match against the computer opponent. Teams received one point for every move they made, two points for every piece they took and three points for each king they earned. Only the scores in the test match were used.

2.1.4. Procedure

After informed consent, each participant was asked to complete three individual tasks: The individual-Raven IQ test; the personality measure; and the mind-in-eyes measure. Subjects were allocated at random into groups of size 2 (12 groups), 3 (8 groups), or 4 (6 groups) – a total of 26 groups. Subjects then completed the five group-IQ tasks. One group did not complete the architecture task due to a procedural error.

2.2. Results

Mean (and SD) for individual Ravens and RME raw scores were 12.23 (2.9) and 26.76 (3.35) respectively. Scores on the individual Ravens and on RME were averaged within each group, and these formed

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

our primary predictors of performance on the group-IQ tasks. Correlations among these group-IQ tasks are shown in Table 1.

A parallel analysis (Horn, 1965) was conducted to determine evidence for the number of factors in the data. This indicated a single general factor be retained (adjusted eigenvalues 1.98, 0.85, 0.69, 0.69, 0.79). This single factor accounted for 39.8% of variance. Testing fit of a 1-factor model using structural equation modeling using *OpenMx* (Neale et al., 2016) and *umx* (Bates, 2014; Bates et al., under review) packages in R (R Core Team, 2016). This indicated that a model with one modification (a covariance between brain storming and group checkers) fit well ($\chi^2(113) = 4.2$, $p = 0.380$; CFI = 0.993; TLI = 0.984; RMSEA = 0.043). In the original report, group-IQ scores generated from the factor analysis using Bartlett's method for deriving factor-scores. Because we had a well-fitting structural model and raw data, we computed latent-factor scores using full-information maximum likelihood modeling (Estabrook & Neale, 2013) to retain information from the group with four rather than five test results. All results were highly similar with both methods.

2.3. Testing causes of group-IQ differences

Our hypotheses regarding the causes of variation in group-IQ scores were tested using a multiple regression with group-IQ score as the dependent variable, and with group size, number of women in the group, along with average individual IQ and average RME scores as predictors. This model accounted for 57% of variance in group-IQ scores, but among the predictors, only individual IQ was significant ($\beta = 0.76$ [0.4, 1.12], $t = 4.37$, $p < 0.001$). Neither proportion-female ($\beta = 0.12$ [−0.2, 0.43], $t = 0.77$, $p = 0.447$) nor reading the mind in the eyes ($\beta = -0.11$ [−0.48, 0.25], $t = -0.65$, $p = 0.520$) were significant. Size of the group approached significance ($\beta = 0.28$ [−0.03, 0.59], $t = 1.87$, $p = 0.075$) suggesting that more people could perhaps accomplish more work. The model could be simplified to one containing only individual IQ as a predictor without significant loss of fit (model comparison $F(3, 24) = 1.562$, $p = 0.23$).

2.4. Discussion

In this first of three studies, we were able to replicate a general factor accounting for over 1/3 of variance in group-IQ test scores. This factor, however, showed strong (rather than weak) loadings on individual IQ. The reported link of group-IQ to numbers of women in the group failed to emerge, and social sensitivity failed to emerge as a significant predictor of group-IQ. This study (the lowest-powered of the three reported here) had 58% power to detect the reported 0.36 effect from social sensitivity to group-IQ (controlling for individual IQ) reported in the largest study of Woolley et al. (2010). The measure of individual IQ (Raven) meant that in study 1, similar test material appeared in both the individual IQ measure and in one of the group tasks. We note that because the "collective intelligence factor" model Woolley et al. (2010) demands a very weak link between group-IQ and individual IQ, and because the group-IQ factor can represent only variance common to all tasks, the collective intelligence factor model predicts such test overlap cannot generate the link we observed between individual cognition and group-IQ. Nevertheless, we thought it desirable to use the Wonderlic IQ test used in the original study 2. For these reasons, we undertook two additional replications, both larger in size (with studies 2 and 3 yielding a combined 95% power to detect a 0.36 effect of social sensitivity on the group-IQ factor), and using the identical individual IQ measure as was used in Woolley et al. (2010) study 2.

3. Replication study 2

In 40 groups, assessed using a total of five group-IQ tasks, we tested the hypotheses that these group tests formed a general group-IQ factor, and that turn-taking, numbers of women, and social sensitivity would be significantly associated with group-IQ scores, contrasting these with models in which individual IQ was the predominant predictor of group scores.

3.1. Method

3.1.1. Subjects

Forty teams of 3 participants were recruited from the public in the city of Chennai, India via contacts made by SG. Sixty-five subjects were male (mean age 26 years, SD 6.1), and 55 were female (mean 26 years, SD 5.8). Subjects were incentivized by a prize of 6000 Indian Rupee raffle, awarded to one group at random after the study completed (to put this in context, a cinema ticket cost ~120 Rupees). An anonymous reviewer requested additional detail on the competence in English language and academic ability of the subjects. English is the official language of India, and fluency in English was a criterion for recruitment. Participants all were educated to college level in English and if in work, the workplace used English as the primary language of communication, as they also interact with international companies. All participants were educated at least to an undergraduate level (or pursuing the same).

3.2. Materials

As in the studies under replication, Reading the Mind in the Eyes (Baron-Cohen et al., 2001) was again used to assess empathizing. To assess individual IQ, we used the Wonderlic (1992) Personnel Test (Form A), as used by Woolley et al. (2010), study 2. This measure consists of 50 multiple-choice items testing spatial, verbal and mathematical ability, with a test time of 12 min. Individuals score one point for each item answered correctly. Scores are highly correlated with those of other intelligence measures, for example, an average correlation of 0.92 with scores on the WAIS has been reported (Wonderlic, 1992).

3.2.1. Group tasks

The five group-tasks were chosen with reference to information on task loadings and correlations on the collective intelligence factor shown in Supplementary Tables S1b and S3b of Woolley et al. (2010). The principal factor loadings reported in Supplementary Table S1b for five tasks common to their study 1 and 2 were high to moderate for three tasks (0.80, 0.72, and 0.61 for Group Matrix Reasoning, Brainstorming (uses of a brick), and Group Typing respectively), lower for

Table 1
Study 1 means, SDs, and correlations for each measure.

	Group Brainstorm	Group Design	Group Shopping	Group Raven	Group Checkers	individual Raven	mean RME
Group Brainstorming	1						
Group building design	0.43	1					
Group Shopping plan	0.42	0.19	1				
Group Raven's IQ	0.37	0.17	0.73	1			
Group Checkers	0.56	0.33	0.37	0.32	1		
Mean individual Raven	0.15	0.07	0.63	0.73	-0.01	1	
Mean individual RME	0.09	0.02	0.22	0.46	-0.14	0.53	1
Mean (SD)	62.96 (12.04)	23.36 (2.39)	48.19 (16.93)	13.96 (2.55)	75.12 (14.21)	12.34 (2.04)	26.78 (2.42)

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

plan-shopping trip (principal factor loading 0.48, correlation with the collective intelligence factor 0.32) and unacceptable for one (principal factor loading of 0.10 for Group Moral Reasoning). Based on these results, we selected the first three tasks and supplemented these with two tasks also used in Woolley et al. (2010) study 2, namely Incomplete Words and Word Completions which showed correlations with the collective intelligence factor of 0.60 and 0.28 respectively (see Woolley et al., 2010, Table S3b). All tasks are described below.

Two tasks assessed Generation. In "brain-storming", groups were asked to generate as many alternate uses of brick as possible within 5 min, with a point scored for each novel use they wrote down. In addition, they were given the verbal fluency task (termed word completions by Woolley et al., 2010), asked to produce as many English words as they could think of that start with the letter *s* and end with the letter *n* within a time limit of 5 min. One point was allocated for each correctly spelled unique word.

To measure "choice", the groups were given the 11-item ICAR Project matrix reasoning test (Condon & Revelle, 2014) as a group-IQ test, with a 10-minute time limit. One point was scored for each item answered correctly. In addition, groups completed the "Incomplete words" task used by Woolley et al. (2010). A set of 36 words with 2–3 letters missing was provided, and the groups were asked to complete as many as possible within 5 min (e.g. "_u_ition" could be correctly completed as "audition").

To measure "negotiation" and "execution", the groups were asked to participate in a Group Typing task, wherein each participant was provided with a hard copy of a difficult text-paragraph. The group then was given 10 min to simultaneously type as much of the text as possible into a shared online Google-docs document. The group was scored based on how percentage-correct of the completed passage at the end of the allotted time.

3.3. Procedure

Participants were met in a receiving room where they gave informed consent, and were assigned at random to a group of 3 people prior to entering the testing environment. A total of 40 groups were tested. Groups were tested in a private room, with facilities supporting testing of up to three groups simultaneously. Participants first completed the individual IQ test, then joined their group and completed the 5 group tasks in a randomized order. A research assistant was assigned to each group, and they administered all 5 tasks to the group. The RA recorded conversational turn taking across the measures to allow a test of the hypothesis that more equal turn taking facilitates group-IQ. In response to a comment from a reviewer, we note that the group interaction was not videoed nor assessed using a proprietary AI-based digital sociometric marker system (as in Woolley et al., 2010 study 1 and 2 respectively), but rather was scored online. During pilot work, we ascertained that in these small groups of three people, taking of turns during the tasks was clearly demarcated, and took place at rates giving ample time for the RA to note down the occurrence of a "turn" in real-time with high reliability. After all tasks were complete, participants completed the Social Sensitivity (RME) task with items presented on computer monitor. Finally, participants were debriefed and given an opportunity to ask the researchers questions regarding the experiment.

3.4. Results

Means, SDs, and correlations for the group and individual measures are shown in Table 2.

To test our first hypothesis that group-IQ test scores would form a general group-IQ factor, we used Horn's parallel analysis, which suggested a 1-factor model of the group-IQ tests (see Fig. 1). A single-factor model also fitted the data well ($\chi^2(5) = 3.14$, $p = 0.678$; CFI = 1.022; TLI = 1.044; RMSEA = 0). For subsequent analyses, scores on this group-IQ factor were extracted from the factor analysis using Bartlett's method. The group-IQ factor accounted for ~50% of group-IQ test variance.

We next tested the role of individual IQ, proportion-female and empathizing using multiple-regression as in study 1. A model with group-IQ as the dependent variable, and average age, average individual IQ, proportion-female and RME as predictors accounted for 85% of variance in group-IQ scores. Individual IQ was again a very strong predictor and highly significant ($\beta = 0.74$ [0.54, 0.94], $F(1, 34) = 57$, $p = 8.6 \times 10^{-09}$). Neither proportion-female ($\beta = -0.04$ [−0.18, 0.1], $F(1, 34) = 0.33$, $p = 0.57$) (see Fig. 2) nor communication ($\beta = -0.07$ [−0.21, 0.07], $F(1, 34) = 1.1$, $p = 0.30$) were significant predictors of group-IQ. Both were in the wrong direction. Unlike in study 1, RME scores were a significant predictor of group-IQ ($\beta = 0.3$ [0.11, 0.5], $F(1, 34) = 9.80$, $p = 0.003$).

3.5. Discussion

Study 2 replicated support for a g-factor among tasks performed by groups, showing also however that this was closely linked to individual group member's IQs. We again found no support for significant effects of number of women or of turn-taking on group-IQ. We did find an association of group-IQ with average RME score. On request from an anonymous reviewer that we acknowledge their thought that (at their request, the comments of the reviewer are not quoted but instead summarized) it is questionable if a raffle could motivate groups to cooperate, we can only state that subjects performed as a group and reported the possibility of a raffle-win to be rewarding. We note also that this post-hoc interaction with payment was not mentioned in the original paper. In our next study subjects are paid, so the theory that monetary reward is required for empathy to cause IQ can be tested there. This anonymous reviewer also suggested that the non-significant effect of turn-taking might increase to significance (which in this study would also require a sign reversal) if turn-taking was scored from video. We disagree, as in groups of three people, the exchange of turns was clearly identifiable, however others might wish to video the sessions and score them offline to test the hypothesis that this significantly alters the data and reveals an otherwise invisible association. Importantly, in the original study, the association of turn-taking with group-IQ did not survive incorporation of empathizing scores in a regression, rendering this question of marginal interest.

Table 2
Study 2 measures: means, SDs, and correlations.

	Group uses	Group MR	Group fluency	Group letters	Group typing	Avg individual Wonderlic	Avg individual RME
Uses of a brick	1						
Matrix reasoning	0.44	1					
Word fluency	0.56	0.59	1				
Missing letters	0.66	0.64	0.8	1			
Group typing	0.31	0.36	0.25	0.28	1		
Individual Wonderlic	0.57	0.75	0.75	0.87	0.49	1	
Avg RME	0.55	0.45	0.60	0.76	0.21	0.69	1
Mean (SD)	12.62 (6.51)	6.62 (2.46)	18.95 (8.1)	62.48 (22.17)	195.05 (68.89)	17.07 (7.4)	20.53 (4.33)

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

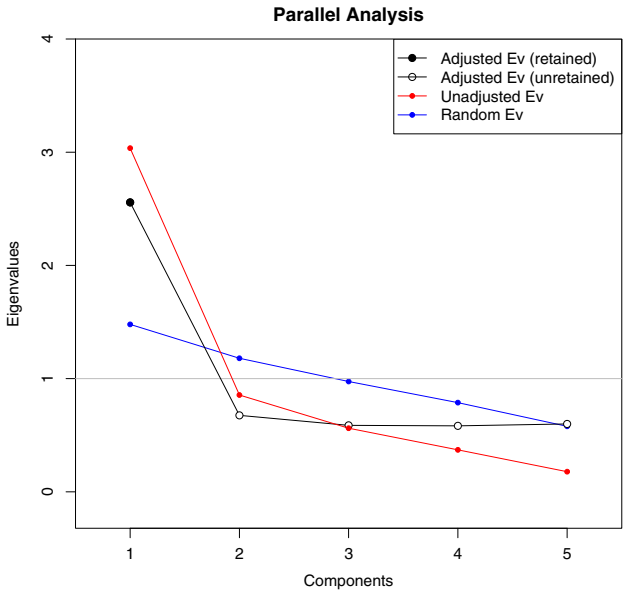

Fig. 1. Parallel analysis of group-IQ tests (study 2).

This reviewer also suggested that (we paraphrase) the mean IQ of this group is far below and variance far above norms for the Wonderlic, and that this caused a stronger than expected correlation of individual-IQ and group-IQ. We considered this argument. The reviewer is hypothesizing that the large effect of individual IQ on group-IQ scores results from low mean and high variance in individual IQ. In evaluating this hypothesis, we would make two points. The mean and SD of Wonderlic scores in the 1992 normative study were 21.06 and 7.12 respectively (Wonderlic, 1992). In the present study, the mean group-average of Wonderlic scores was 17.07 and SD was 7.4. It is false, therefore, to say that the standard deviation was far above that of the normative sample: Rather than being far above the normative SD, the sample SD was highly similar. Thus there is no range effect to correct. Moreover, the reviewers' account of the large effect we observed of individual on group-IQ entails hypothesizing that, as we find, these variables do indeed correlate in the population, and contrary to the prediction from group-IQ theory, that this accounts for much of the variance in group-IQ.

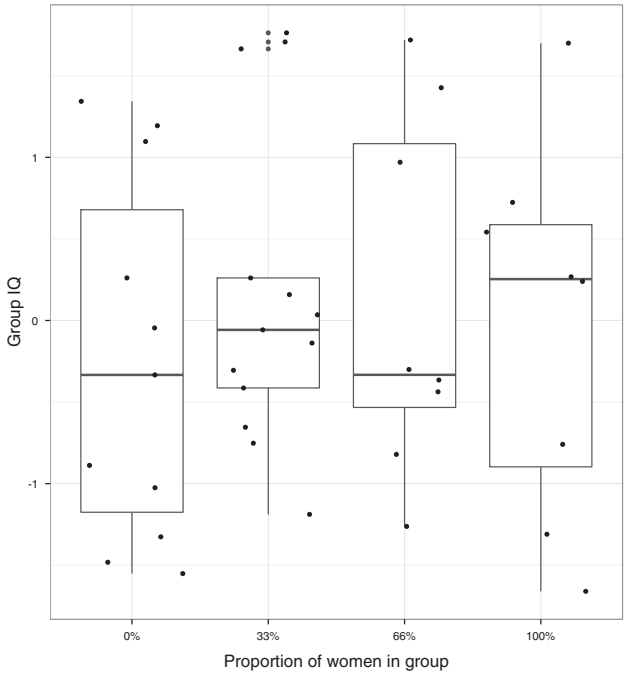

Fig. 2. Group-IQ as a function of number of women in the group (NS).

Turning to the claim that the present finding is caused by a lower mean Wonderlic score in this study compared to that of the Wonderlic normative sample, we highlight two considerations. First, unlike variances, mean differences have no effect on the value of a correlation. Thus a mean difference cannot induce a correlation between individual and group IQ. Second, we note that this comment from the reviewer implies an additional hypothesis, namely that individual IQ and group-IQ are related, but more strongly at lower levels of individual IQ than at high levels of individual IQ. There is no support for this hypothesis in the data, but the apparent critique predicts that organizations seeking high group-IQ should select strongly on individual IQ to avoid low group-IQ, which, again, contradicts the group-IQ theory prediction that individual and group-IQ are largely independent.

To increase our power to model effects among these variables, we next completed a third replication study identical to study 2, but conducted in the UK instead of in India.

4. Replication study 3

In study three, we sought to gather further evidence on the relationship of individual and group-IQ and to better understand the mechanism of this association. Note: in addition to replicating the identical suite of group-IQ tasks used in study 2, we added a questionnaire (the moral foundations questionnaire (Graham et al., 2009) and an experimental manipulation (conducted after the study was completed). These were constructed to allow us to explore the claim that "*groups that had smart people dominating the conversation were not very intelligent groups*" (Woolley & Malone, 2011, p. 2). After completing the replication study, groups were allocated at random to one of three conditions: authority, empathizing, or control. In the authority condition, the subject with the highest WPT score was selected to be group leader, and the rest of the group was instructed to allow this person to direct problem-solving and control decisions. In the empathizing condition, group members were instructed to ensure each person had an equal amount of talking time and to pay attention not only to what group members were saying, but how they were saying it. These interventions were prompted both by the novel result in study 2, and to test whether groups might seek to raise their IQ performance by adopting new habits. Whereas in study 2 we saw that communication was unrelated to group-IQ, here we contrast explicit promotion of an individual to coordinate group activity testing if this lowers group-IQ, and contrasting this with an explicit turn-taking manipulation, testing if this raises group-IQ.

4.1. Method

4.1.1. Subjects

Forty teams of three participants were recruited from the general public in the city of Edinburgh, Scotland. Forty-four were male (76 Female) and ages ranged from 17 to 63 years (Mean = 24.23, SD = 9.01).

4.1.2. Materials

All materials were identical to that of study 2 with the addition of a final group-IQ measure taken after a manipulation encouraging either authority or empathy. For this purpose, the final set of the Raven's Standard Progressive Matrices (SPM, Set E) was used. This test is a standardized measure of fluid intelligence (Raven, Raven, & Court, 1998). Groups were asked to complete the test as quickly and accurately as possible, and were given a 10-minute time limit. The number correct and time taken to complete all items was recorded for each group.

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

4.2. Procedure

Procedures were identical to those of Study 2: Subjects were welcomed into the testing room, where they completed the ethics. Next, subjects completed the same Wonderlic individual intelligence test, and RME. They were then formed into groups and completed the same 5 group-IQ tasks as were used in Study 2 to assess group intelligence. Group compositions by sex were 1, 14, 13, and 12 for no women, one, two, and three women in the group respectively. The RAs remained in the testing room, but as part of their instruction, participants were told to ask all questions before the task was started and before we started timing. During the task, no assistance was provided.

Because we had hoped that group personality-linked differences would emerge, an experimental manipulation was planned. Though no group-IQ differences emerged which were not well modeled by individual IQ, this manipulation which occurred after the main study was complete is recorded here for completeness. After the group-IQ replication was complete, groups were randomly assigned to either control ($n = 14$ groups), empathy ($n = 14$ groups) or authority ($n = 12$ groups) manipulation. Both experimental groups were told "*we are testing a new strategy that has proven to enhance performance in previous studies*". In the empathy-inducing manipulation, the instructions emphasized the role of emotional understanding and empathy toward one another. These groups were asked to ensure that each person in the group received an equal amount of talking-time, that no person was to dominate the group and, lastly, to pay attention not only to what their group members were saying, but how they were saying it, that is to focus on one another's body language, facial expressions and tone.

In the Authority manipulation condition, subjects were told that "*evidence showed the best leadership strategy was one in which one individual shoulders the leading role*" and that a leader would be chosen based on their ability at the IQ task. The person with the best WPT score was appointed group leader. Groups were asked to allow the leader to direct problem-solving and make the ultimate decisions. They were asked to try and work cohesively under the assigned authority. No intervention/strategy was provided to the control group. They simply were asked to undertake a final task together. After instruction, groups in each condition then completed the Raven's Standard Progressive Matrices (SPM, Set E) items.

4.3. Results

The test scores again showed the positive manifold characteristic of IQ (see Table 3) and a parallel analysis again indicated a single factor accounted for the score data. Scores on this factor were again computed for each group.

As in study 2, we tested if communication, number of women in the group, or emotional empathizing were associated with enhanced group-IQ using linear models with group-IQ scores as the DV. To avoid any suppression of these variables' effects, only age was covaried. No significant effects were found for communication ($\beta = 0.01$ [−0.37, 0.40], $t = 0.06$, $p = 0.95$). Neither was any significant effect of number of women in the group on group-IQ. To further explore links of group-IQ to gender makeup, we tested linear ($\beta = 0.18$ [−0.15, 0.52], $t = 1.11$, $p = 0.27$) and quadratic effects ($\beta = -0.033$ [−0.42, 0.35], $t = -0.17$, $p = 0.86$) but none were significant.

We next tested the effect of RME. In a simple linear model predicting group-IQ, and controlling only age, mind in the eyes reached significance ($\beta = 0.37$ [0.05, 0.70], $t = 2.34$, $p = 0.025$). We then tested the predicted independence of this effect from individual IQ by adding individual IQ to the model. The effect of individual IQ was large and highly-significant ($\beta = 0.67$ [0.42, 0.92], $t = 5.52$, $p < 0.001$). Contrary to prediction, adding individual IQ had the effect of rendering RME non-significant ($\beta = 0.15$ [−0.10, 0.41], $t = 1.22$, $p = 0.23$).

Finally, before moving to combine the data from study 2 and study 3, we analyzed our attempted experimental manipulation of group-IQ via instructions to the groups to either obey the brightest person in the group, treating them as an authoritative leader (authority condition), to attend to each other, ensuring that all subjects were listened too (empathizing), and comparing these conditions to a control group. To test this manipulation, we ran a multiple regression, predicting group-Ravens score (the new group-IQ test) with average age, IQ, and empathizing as well as the manipulation as predictors. There was no evidence for any effect of the authority/empathy manipulation ($F(2, 34) = 1.14$, $p = 0.33$). The standardized (β) effects of the Authority and Empathy conditions relative to the control condition were in fact both negative (i.e., worse: −0.42 (95% CI[−0.98, 0.15], $t = -1.5$, $p = 0.142$) and −0.13 (95% CI[−0.68, 0.41], $t = -0.5$, $p = 0.620$ respectively). We do not, however, place to much emphasis on this result for the following reason. Our choice of Raven items allowed several of the groups to reach ceiling-level scores, suppressing group-IQ variance. Interestingly under these conditions, average IQ also only "approached significance" ($F(1, 34) = 2.92$, $p = 0.097$), indicating that one factor which might suppress effects on group-IQ is ceiling or floor effects.

We next moved to combine the datasets from studies 2 and 3, and to generate a model which best accounts for the roles of RME and IQ on group-IQ.

4.4. Joint structural modeling study 2 and study 3 data

Because studies 2 and 3 used identical test materials and methods, we were able to combine these into a single data set, controlling for study, to gain power and to use structural equation modeling to directly compare competing models of the causes of group-IQ. We note that an anonymous reviewer suggested that the studies should be left separate rather than be combined. All the joint analyses reported below controlled for study origin, either in the regression analyses (as a covariate to account for differences between the study populations) or (in the case of the structural modeling) by regressing study out of the raw data prior to modeling. As in all previous analyses, we controlled for average age.

The study covariate showed only very small, non-significant effects, validating the combination of data from the two sites. Importantly, the broad findings for each study individually replicated in the joint data. In the joint data, number of females in the group was not a significant predictor in a model controlling for study and age ($F(3.74) = 0.99$,

Table 3
Study 3 means, SDs, and correlations for the group and individual measures.

	Group uses	Group Raven	Group fluency	Group letters	Group typing	Avg individual Wonderlic	Avg individual RME
Uses of a brick	1						
Group Ravens	0.18	1					
Word fluency	0.39	0.12	1				
Missing letters	0.17	0.24	0.61	1			
Group typing	0.19	0.14	0.06	0.02	1		
Avg individual Wonderlic	0.32	0.38	0.71	0.54	-0.04	1	
Avg RME	0.12	-0.15	0.34	0.28	-0.16	0.33	1
Variable mean (SD)	21.05 (5.52)	7.5 (1.8)	24.8 (6.78)	83.58 (10.82)	23.01 (2.84)	25.46 (3.9)	26.55 (2.56)

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

$p = 0.40$). Neither were any of the linear, quadratic and cubic functions of number of females significant ($p = 0.29$, 0.12, & 0.72 respectively).

In a regression with group-IQ as the DV and study, average age, average IQ, and average RME as predictors, RME appeared to have a significant effect on group-IQ scores ($\beta = 0.191$ [0.032, 0.349], $t = 2.401$, $p = 0.019$) Average individual IQ alone showed a powerful effect (on its own, controlling for study and age, accounting for 81% of variance in group-IQ (standardized beta 0.86 [0.74, 0.99] ($t = 13.6$ $p < 2 \times 10^{-16}$: see Fig. 3).

If these linear models alone were used to guide our conclusions, it would be reasonable to conclude that numbers of women and turn taking do not affect group-IQ, that individual IQ plays a very strong role, but that RME does appear also to play a role, albeit much smaller than predicted. However, a more powerful method is available to test complex competing models of the relationship of individual predictors of group-IQ and the latent variable they form with individual IQ and empathizing, and we next turned to structural equation modeling to formally test 4 alternative models.

4.4.1. A structural model of group-IQ

A key advantage of the SEM framework for testing models is incorporating a measurement model for latent traits such as group-IQ. This allowed us to test not only how our predictors jointly affect each other, but crucially allowed testing whether RME is associated with the latent group-IQ factor, or with specific group-test variance, distinct from this factor. A specific recent example using SEM for this purpose can be seen in a test of the hypothesis that education raises general ability or acts directly on individual school subjects (Ritchie, Bates, Der, Starr, and Deary, 2013). We next outline the 4 distinct theoretical models for the data, taking advantage of this latent-variable SEM approach.

4.5. Four alternative models of the origins of group-IQ performance

The alternative models we tested are shown in Fig. 4. In model **A**, group-IQ is proposed to result from individual differences in empathizing. In model **B**, only individual IQ affects group-IQ variance. Model **C** combines both effects, with group-IQ modeled as reflecting both empathizing and individual IQ, with empathizing also loading on individual IQ, reflecting the association of these two traits. This model is the one perhaps suggested by multiple regression approaches, with both individual IQ and individual RME traits acting to raise group-IQ. Finally, model **D** suggests a very different causal situation. In this model, only average individual IQ affects group-IQ. Empathizing has no effect on the latent group-IQ factor, but is allowed to affect one or more single test scores. Each of these models was built and compared using *OpenMx* (Neale et al., 2016) and *umx* (Bates, 2014; Bates et al., under review) packages in R (R Core Team, 2016).

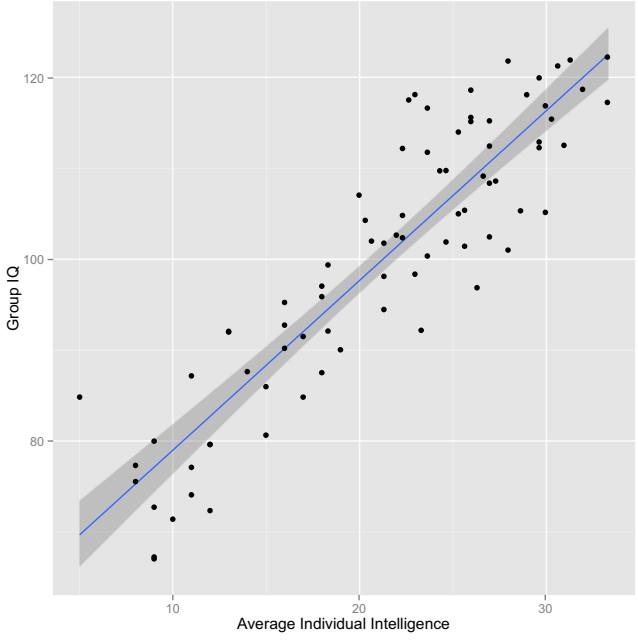

Fig. 3. Relationship of individual IQ to group-IQ in the combined data from studies 2 & 3.

The results of these comparison models were as follows. Model **A**, in which mind in the eyes was used to account for group-IQ fit poorly ($\chi^2(537) = 105.65$, $p < 0.001$; CFI = 0.601; TLI = 0.301; RMSEA = 0.312), and significantly worse than any other model tested. Viewing Group-IQ as emerging from empathizing alone, then, provided a very poor fit to the data. Modeling group-IQ entirely as consequence of variance in individual IQ (model **B**) fit significantly better than model A (AIC 2237.0 vs 2320.1), but did not reach modern standards of good fit ($\chi^2(537) = 22.52$, $p = 0.032$; CFI = 0.955; TLI = 0.921; RMSEA = 0.105). Modeling group-IQ as an outcome of both IQ and mind in the eyes (model **C**) also failed to generate an acceptable fit ($\chi^2(536) = 20.14$, $p = 0.043$; CFI = 0.961; TLI = 0.926; RMSEA = 0.102). This model also did not fit better than model B ($\chi^2(1) = 2.39$, $p = 0.12$).

4.6. Best fitting model: no effect of empathizing on the group-IQ factor

The best fitting model was model **D**, in which empathizing was constrained to have no impact on the latent group-IQ factor, and instead was allowed only to co-vary with individual IQ and to influence single group-IQ tasks directly (as opposed to being mediated via the group-IQ latent trait. We could not determine in advance which traits RME might affect, and so tested a model in which RME was allowed to load on all the group-IQ measures. This model did not fit significantly better than the model B (Only IQ influencing group-IQ) ($\chi^2(5) = 7.12$), $p = 0.21$), suggesting that not all paths from RME to specific manifests had appreciably improved model fit. We next sought, therefore to remove the unnecessary paths. Inspecting the fitted model showed clearly that paths from RME to all but the missing-letters task were small, and these were dropped without significant loss of fit ($\chi^2(4) = 2.76$, $p = 0.60$) and with an improved AIC which decreased from 2239.8 to 2234.6. This new model fit better than model B ($\chi^2(1) = 4.36$, $p = 0.037$) and became our best candidate, with better fit than all alternative models ($\chi^2(536) = 18.16$, $p = 0.078$; CFI = 0.969; TLI = 0.942; RMSEA = 0.09). Preparing a final best-fitting model, we examined modification indices. These suggested three manifest covariances might be added, and this yielded a well-fitting model by modern criteria ($\chi^2(533) = 9$, $p = 0.342$; CFI = 0.996; TLI = 0.989; RMSEA = 0.039). This best-fitting model is shown in Fig. 5. The added covariances are in-line with contemporary hierarchical intelligence test theory (Carroll, 1993) in which group factors lie beneath general ability. These should cause some quadrants or group-IQ to associate more with each other, as observed. The model estimated the residual variance of group-IQ as zero after accounting for individual IQ (i.e., all paths shown in Fig. 5 are free and show their maximum likelihood values). This meant that group-IQ could in fact be modeled as completely determined by average individual (residual variance set to 0.0 and the path from average individual IQ to group-IQ fixed at 1.0) with no change in fit ($\chi^2(2) = 0$, $p = 1.000$) and yielding an economical model with high-fidelity to the data ($\chi^2(535) = 9$, $p = 0.53$; CFI = 1.00; TLI = 1.00; RMSEA = 0).

5. General discussion

The three studies reported here and, especially, the joint modeling cast important light on the origins of high cognitive performance in groups. Rather than a small link of individual IQ to group-IQ, we found that the overlap of these two traits was indistinguishable from 100%. Smart groups are (simply) groups of smart people. By contrast, we

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

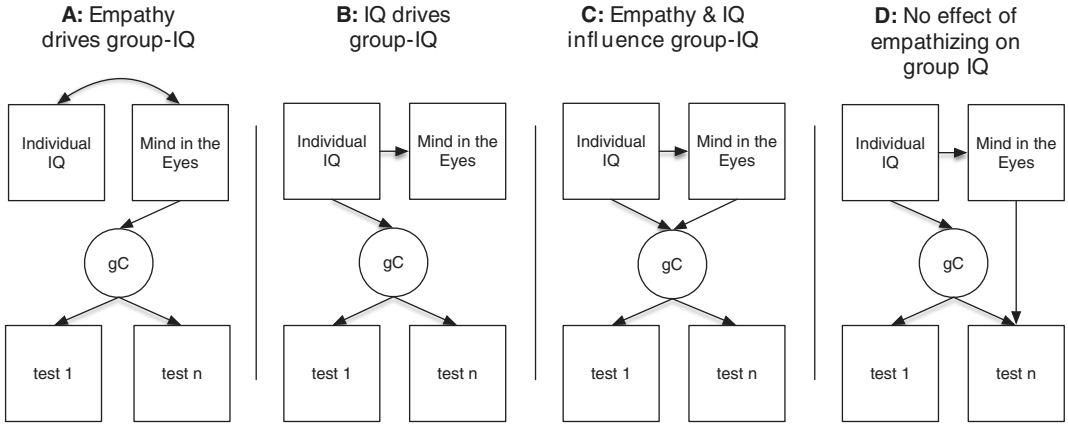

Fig. 4. Four models contrasted. **A**: group-IQ results from differences in empathizing. **B**: only individual IQ affects group-IQ. **C**: group-IQ reflects both empathizing and individual IQ. **D**: group-IQ reflects of individual IQ. Empathizing is also part-dependent on IQ, but has zero effect on group-IQ.

found little to no evidence for two proposed causes of group-IQ: numbers of women in the group and turn-taking, and found evidence for a weak and specific impact of RME on one group task, but not on latent group-IQ. These findings are elaborated on below.

The finding that IQ and group-IQ can be set equal bolsters studies reported in work-performance showing that groups of bright individuals outperform groups of less able individuals (Devine & Philips, 2001). We take this work to a new level, suggesting that, in terms of latent group-IQ, group performance reflects nothing beyond individual contributions to average IQ. Thus we found no support for the hypothesis that "*group intelligence [has] relatively little to do with individual intelligence*" (Woolley & Malone, 2011, p. 2).

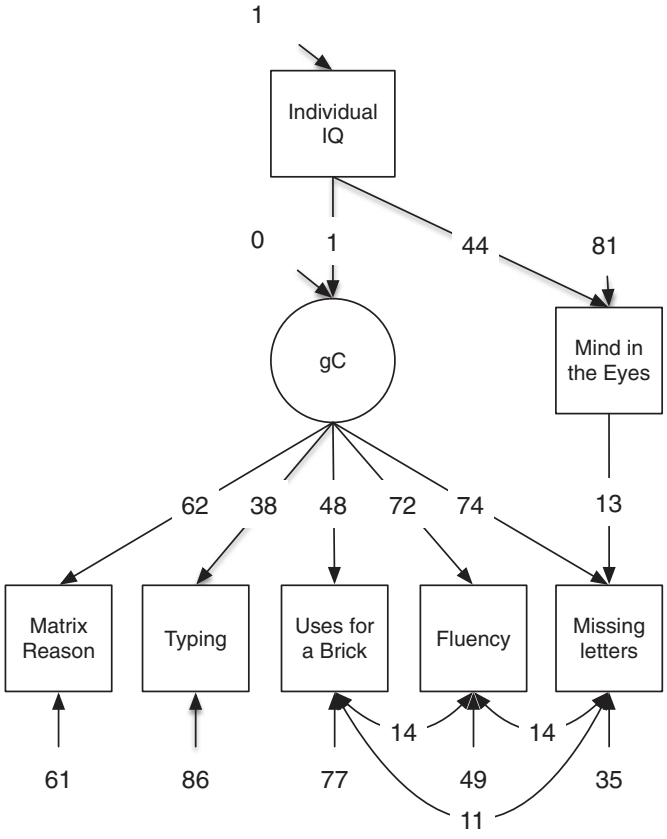

Fig. 5. Best fitting model of combined Study 2 and 3 data.

We were able to conduct a direct test of the causes of group-IQ examining if this factor "*appears to depend both on the composition of the group (e.g., average member intelligence) and on factors that emerge from the way group members interact when they are assembled (e.g., their conversational turn- taking behavior)*" Woolley et al. (2010, p. 688). Across the three studies we saw no significant support for the hypothesized effects of women raising (or men lowering) group-IQ: All male, all female and mixed-sex groups performed equally well. Nor did we see any relationship of some members speaking more than others on either higher or lower group-IQ. These findings were weak in the initial reports, failing to survive incorporation of covariates. We attribute these to false positives. The equal performance of groups irrespective of gender is in-line with previous findings that men and women have near-identical mean IQs (though males have greater variance) (Deary, Irwing, Der, & Bates, 2007), and the strong dependency reported here of group-IQ on individual IQ. The present findings cast important doubt on any policy-style conclusions regarding gender composition changes cast as raising cognitive-efficiency.

5.1. Comparing the present results to those previously reported

Our multiple regression results in study 2 (but not in study 1 or study 3) yielded an apparent correlation between group-IQ scores and average empathizing scores as measured by the RME. In an innovation not used in the original reports, we tested this relationship using an SEM approach. Unlike regression, this was able to discriminate specific-test effects from an association with the group-IQ latent trait. Translating the data into this SEM framework gave a very different interpretation of the link of group-IQ with empathizing. Model comparisons revealed that empathizing performed inadequately (Model **A**) as an explanation of variance in group-IQ. In our preferred model (see Fig. 5), RME effects on group-IQ tasks were reduced to a single, low magnitude, test-specific effect on the missing-letter generation task. It might be that the task of deciding which missing letters complete a valid word involved components of cognition/collaboration specific to this test, and tapped by RME.

In terms of outcome discrepancies, the original report of a modest apparent role for empathizing in group-IQ can be accounted for within the model supported here. First, the reported social underpinnings of group-IQ are far from complete: correlations with group-IQ of only around $r = 0.2$ were reported for proportion of females and turn-taking variables, and ~0.3 for social sensitivity Woolley et al. (2010). Importantly, only the link to social sensitivity was significant in their full models. Thus what is in question on the social hypothesis of group-IQ is around 10% of shared variance between group-IQ and social sensitivity. As reflected in our data, to the extent that social intelligence is an intelligence (Locke, 2005), it is both correlated with individual general IQ,

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

and contains some trait-specific variance. A parsimonious explanation of why regression models might show social sensitivity correlating with group-IQ, then, is captured in our final model (see Fig. 5) – social sensitivity is itself modestly associated with high personal IQ. A specific linkage of sensitivity to a single group task that weights this more heavily is also represented, with no impact of social sensitivity to group-IQ.

Less clear is why we found a strong link from individual average IQ to group-IQ scores across three independent studies, reaching identity in our combined model, when the original reports estimated this association at <0.3. We attribute this to two factors. The first is the use of structural modeling that captures and represents covariation among multiple variables which regressions cannot. If this were the only factor, however, previous researchers should also have found a heavy dependence of group-IQ on individual IQ. We can only attribute the discrepancy to some factor limiting the validity or range of the IQ measures in Woolley et al. (2010) (the only other report (Engel et al., 2014) did not measure individual IQ). Ceiling and/or floor effects in testing, for instance, could suppress the link between individual IQs and group-IQ. Accepting this model also seems more parsimonious given the by now very wide validation of the biopsychosocial development and mechanisms of individual IQ (Bates, Lewis, & Weiss, 2013; Hill et al., 2014) as well as its general nature, and strong links to performance: Individual IQ appears, therefore, as an adequate account of novel problem solving in groups (as opposed to longer-term cooperative implementation of such novel ideas, which likely involves cooperation and conscientiousness).

5.2. Limitations and future directions

5.2.1. Power

A reviewer argued that, despite our combined studies all showing the correlation among group tasks and detecting effects of cognitive ability predicted to be weak, the studies suffer unacceptably low power and were unsuitable for publication given conventional standards for high power in peer-review. We respond to this claim in three parts. First, it is, sadly, not the case that power is uniformly above 80%. For instance Button et al. (2013, p. 365) found "*the average statistical power of studies in the neurosciences is very low*". Second, study 1 of the paper we are attempting to replicate (Woolley et al., 2010), published in *Science*, used $n = 40$ groups, as did our study 2. All of this would not matter, however, if we lacked power to detect the effects under study – power, indeed matters. We therefore address this question in more detail here.

Importantly, in this set of three studies, we are not testing against a so-called null model, where a claimed effect is simply not found to be significant. Instead, we are able to compare two competing models: That proposed by Woolley et al. (2010) which predicts that individual IQ has a negligible impact on group-IQ, while empathizing has a large effect, and the competing model which we develop here based on the outcome of Study one. In our model, individual IQ has a large effect and empathizing does not impact on general ability. We can thus compare model fits as well as compare significance of individual parameters. A lack of power would lead to a lack of discriminability between the two models, but, in a world where empathizing has a strong effect, would favor recovery of that model across the studies. However, across all three studies we found significantly better fit for the individual-IQ model of group-IQ versus the empathizing or sex or turn taking models. We were reliably able to detect effects of individual IQ which were predicted to be vanishingly small, while simultaneously seeing estimates of empathizing effects, claimed to be much larger than those of individual IQ, estimated as most likely at or near zero. Confidence in the results is further buttressed by our replication of the predicted correlation among individual group-IQ tasks: The general factor emerging from these was detected in all three studies, and all group-IQ tests loaded significantly on the factor. Finally, the power in the combination of study 2 and 3 to detect the effects proposed for empathizing exceeded 95%. We therefore find the claim that we lacked power to detect a predicted large effect to be internally inconsistent with our reliable detection of effects predicted to be smaller, and not compatible with either the single study outcomes or the three results taken jointly.

5.2.2. Did we, in fact, undertake a replication?

A reviewer concluded that we (paraphrasing) need to… remove any claim that [we] undertook a replication. Regarding whether we have undertaken a replication, we direct readers to our use of the IQ tasks used in either Woolley et al. (2010) study 1 (Matrix reasoning) or study 2 (Wonderlic), our use of the same measure of empathizing (Baron-Cohen et al., 2001), and our use of group-IQ tasks selected from among those used in Woolley et al. (2010) as performing *best* in that work (see above, study 2, methods for task-choice rationale). The reviewer suggested that our choice of tasks shown by Woolley et al. (2010) to be the best measures of collective IQ in fact caused our failure to find any role for empathizing in group IQ. We see no logical mechanism for such as effect: picking good measures is instead a strength of this study. It is the case that we were unable to replicate the measurement of turn-taking in Woolley et al., 2010 study 2 which recorded $n = 47$ groups using sociometric badge technology and an in-house proprietary algorithm identifying individual speakers in sound streams digitized from the badges, then identifying algorithmically when an individual stopped speaking while another started, then began speaking to segment the multiple sounds streams and, ultimately tabulate turn taking. Instead, we identified turns taken by each of the three group members by marking down when each member took a turn speaking. In these small groups marking down each turn taken proved reliable and relatively effortless. This difference of counting how many turns were taken live, rather than from recorded tape, prompted a reviewer to insist we delete all analyses of conversational turn-taking (the same reviewer requested us in fact to delete all of study 2). Perhaps future researchers will experiment with these recording methods and generate a measure of their inter-correlation. Here we simply highlight for the reader that, in the Woolley et al. (2010) studies, turn-taking provided no independent prediction of group-IQ, rendering the point moot. Given that in the original report turn-taking was viewed as an outcome of social sensitivity (which rendered turn-taking effects non-significant), it is unclear if improved assessment of this proximal variable is warranted given the lack of effect of social sensitivity.

5.2.3. Unidentified moderators

An anonymous reviewer suggested that (paraphrasing) there clearly must be an unidentified moderator which accounts for why individual IQ and collective intelligence correlated so strongly. Readers should evaluate this claim for themselves. It is far from clear to us that an unidentified moderator "must" exist. As noted above, the original report showed an association of individual IQ with group-IQ, the association of empathizing reported was comparable to that found here, and the reported effects of sex, and turn-taking were non-significant in models with suitable covariates. As discussed above, moderators such as mean or variance in individual IQ, or differences in culture cannot account for the effects found here without substantively modifying the original theory: For instance, to claim that there is a strong link of individual IQ to group-IQ at IQ levels around 90, but not at IQs of 100. Critically, such post-hoc modifications cannot be invoked consistently across the three studies without mutual contradiction, e.g. with the replication result from the UK. Instead we suggest the most cogent analysis is that the proposed moderators were not in fact operative and the data show consistent effects across study (e.g. Fig. 3).

5.2.4. More general considerations for future study of IQ in groups and organizations

Human performance in groups has been a topic of the highest interest (Simon, 1997) over a long period of time (Bouchard, 1969). Alongside models of cognitive ability (Deary, Spinath, & Bates, 2006),

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

cooperation and competition (Lewis & Bates, 2010, 2011, 2013, 2014), and systems for cumulative culture (Dean, Kendal, Schapiro, Thierry, & Laland, 2012), the solving of novel problems in groups must rank among the most practically important of topics available in psychology (Wechsler, 1971). Much of course has yet to be learned. Here we mention just two possible directions for extending work on the role of cognitive ability on problem solving in groups.

It is important here to distinguish very broad concepts such as "successful teams", judged by long-term implementation of agreed goals, from specific constructs such as group-IQ. The group sizes used in group-IQ research are very common in small teams. They are nevertheless much smaller than those assembled in human innovations such as companies, cities, nations, and armed forces. The periods of time involved are also brief, and the outcome measure (novel problem solving) is only one component of organizational success. Organizations universally involve complex sets of norms and rules, working toward agreed goals, and extended lifespans. It would be valuable to extend studies of IQ to examine performance in these much larger groupings, and involving contributing to a group goal or norm-maintenance (where altruism and agreeableness may be important: Lewis & Bates, 2011, 2014), or over longer periods of time (where, for instance individual conscientiousness may be relevant for extended or multi-stage tasks: Jackson et al., 2010; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007).

It is interesting also that groups did not perform better than individuals – a genuine group-IQ might be expected to enable problem solving to scale linearly (or better) with number of subjects. In group-IQ tasks, coordination costs appear to prevent group problem-solving from rising even to the level of a single individual's ability. This implicates not only unsolved coordination problems, which are well-known barriers to scale (Simon, 1997) but also reiterates the finding that the individual problem-solver remains the critical reservoir of creativity and novel problem solution (Shockley, 1957).

5.3. Conclusion

In conclusion, across three studies groups exhibited a robust cognitive g-factor across diverse tasks. As in individuals, this g-factor accounted for approximately 50% of variance in cognition (Spearman, 1904). In structural tests, this group-IQ factor was indistinguishable from average individual IQ, and social sensitivity exerted no effects via latent group-IQ. Considering the present findings, work directed at developing group-IQ tests to predict team effectiveness would be redundant given the extremely high utility, reliability, validity for this task shown by individual IQ tests. Work seeking to raise group-IQ, like research to raise individual IQ might find this task achievable at a task-specific level (Ritchie et al., 2013; Ritchie, Bates, & Plomin, 2015), but less amenable to general change than some have anticipated. Our attempt to manipulate scores suggested that such interventions may even *decrease* group performance. Instead, work understanding the developmental conditions which maximize expression of individual IQ (Bates et al., 2013) as well as on personality and cultural traits supporting cooperation and cumulation in groups should remain a priority if we are to understand and develop cognitive ability. The present experiments thus provide new evidence for a central, positive role of individual IQ in enhanced group-IQ.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The reading the mind in the eyes test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Bates, T. C. (2014). umx: A package for SEM in R (Version.9) [R package]. Edinburgh, UK github.com

Bates, T. C., Lewis, G. J., & Weiss, A. (2013). Childhood socioeconomic status amplifies genetic effects on adult intelligence. *Psychological Science*, *24*(10), 2111–2116.

Bates, T. C., Neale, M. C., & Maes, H. H. (2016). umx: A library for structural equation and twin modelling in R. *Journal of Statistical Software* (under review).

Bouchard, T. J. (1969). Personality Problem-Solving Procedure and Performance in Small Groups. *Journal of Applied Psychology*, *53*(1p2), 1–29.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY, USA: Cambridge University Press.

Christie, R., & Geis, F. (1970). *Studies in Machiavellianism.* NY: Academic Press.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64.

Costa, P. T., Jr., & McCrea, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

Dean, L. G., Kendal, R. L., Schapiro, S. J., Thierry, B., & Laland, K. N. (2012). Identification of the social and cognitive processes underlying human cumulative culture. *Science*, *335*(6072), 1114–1118.

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother-sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY 1979. *Intelligence*, *35*(5), 451–456.

Deary, I. J., Spinath, F. M., & Bates, T. C. (2006). Genetics of intelligence. *European Journal of Human Genetics*, *14*(6), 690–700.

Devine, D. J., & Philips, J. L. (2001). Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research*, *32*, 507.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PloS One*, *9*(12), e115212.

Estabrook, R., & Neale, M. (2013). A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. *Multivariate Behavioral Research*, *48*(1), 1–27.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002.

Hill, W. D., Davies, G., van de Lagemaat, L. N., Christoforou, A., Marioni, R. E., Fernandes, C. P. D., et al. (2014). Human cognitive ability is influenced by genetic variation in components of postsynaptic signalling complexes assembled by NMDA receptors and MAGUK proteins. *Translational Psychiatry*, *4*, e341.

Horn, J. L. (1965). A rationale and a test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511.

Lewis, G. J., & Bates, T. C. (2010). Genetic evidence for multiple biological mechanisms underlying ingroup favoritism. *Psychological Science*, *21*(11), 1623–1628.

Lewis, G. J., & Bates, T. C. (2011). A common heritable factor influences prosocial obligations across multiple domains. *Biology Letters*, *7*(4), 567–570.

Lewis, G. J., & Bates, T. C. (2013). Common genetic influences underpin religiosity, community integration, and existential uncertainty. *Journal of Research in Personality*, *47*(4), 398–405.

Lewis, G. J., & Bates, T. C. (2014). Common heritable effects underpin concerns over norm maintenance and in-group favoritism: Evidence from genetic analyses of right-wing authoritarianism and traditionalism. *Journal of Personality*, *82*(4), 297–309.

Locke, E. A. (2005). Why emotional intelligence is an invalid concept. *Journal of Organizational Behavior*, *26*(4), 425–431.

McGrath, J. E. (1984). *Groups: Interaction and performance.* Englewood Cliffs, NJ: Prentice-Hall.

Nature Editorial (2016). Go forth and replicate! *Nature*, *536*(7617), 373.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., et al. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549.

R Core Team (2016). *R: A language and environment for statistical computing (version 3.3.0).* Vienna, Austria: R Foundation for Statistical Computing.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's advanced progressive matrices and vocabulary scales* (1998 ed.). Oxford, England: Oxford Psychologists Press Ltd.

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308.

Ritchie, S. J., Bates, T. C., Der, G., Starr, J. M., & Deary, I. J. (2013). Education is associated with higher later life IQ scores, but not with faster cognitive processing speed. *Psychology and Aging*, *28*(2), 515–521.

Ritchie, S. J., Bates, T. C., & Plomin, R. (2015). Does learning to read improve intelligence? A longitudinal multivariate analysis in identical twins from age 7 to 16. *Child Development*, *86*(1), 23–36.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345.

Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. *Proceedings of the IRE*, *45*(3), 279–290.

Simon, H. A. (1997). *Administrative behavior: A study of decision-making processes in administrative organizations* (4 ed.). New York, USA: Macmillan.

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–293.

Wechsler, D. (1971). Concept of Collective Intelligence. *American Psychologist*, *26*(10), 904–907.

Wilson, R. C., Guilford, J., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, *50*(5), 362.

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004

Wonderlic, E. F. (1992). *Manual of the Wonderlic personnel test.* Libertyville: E.F. Wonderlic & Associates, Inc.

Wonderlic, E. F., & Hovland, C. I. (1939). The personnel test: A restandardized abridgment of the Otis SA test for business and industrial use. *Journal of Applied Psychology*, *23*(6), 685–702.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Woolley, A. W., & Malone, T. (2011). What makes a team smarter? More women. *Harvard Business Review*, *89*(6), 32–33.

Woolley, A. W., Malone, T. W., & Chabris, C. F. (2015). Why some teams are smarter than others. *New York Times*. from http://www.nytimes.com/2015/01/18/opinion/sunday/why-some-teams-are-smarter-than-others.html

Please cite this article as: Bates, T.C., & Gupta, S., Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups, *Intelligence* (2016), http://dx.doi.org/10.1016/j.intell.2016.11.004