MIT Sloan School Working Paper 5431-18

Measuring Collective Intelligence in Groups: A Reply to Credé and Howardson

Anita Williams Woolley, Yeonjeong Kim, Thomas W. Malone

Measuring Collective Intelligence in Groups: A Reply to Credé and Howardson

Anita Williams Woolley[a], Yeonjeong Kim[b], Thomas W. Malone[b]

[a] Tepper School of Business, Carnegie Mellon University
awoolley@cmu.edu

[b] MIT Sloan School of Management
{yeonkim, malone}@mit.edu

Abstract. Recent work by Woolley, Chabris, Pentland, Hashmi, & Malone (2010) and their colleagues finds evidence for a general collective intelligence factor that predicts a group's performance on a wide variety of tasks, like the general intelligence factor does for individuals. Credé and Howardson (2017) argue that there is not yet sufficient evidence to conclude such a collective intelligence factor exists. Specifically, C&H suggest that the general factor is not strongly enough correlated with all the tasks examined, but we point out problems with their interpretation of the evidence presented. C&H also suggest that the effort and abilities of the group members are "statistical artifacts" that may have inflated the correlations observed, but we demonstrate that the data do not support that interpretation. We concur with C&H about the importance of improving measures of collective intelligence, and we elaborate upon their suggestions for future research to better understand the boundaries and the causal mechanisms of this phenomenon.

Keywords. collective intelligence, group performance, construct validity

1. Introduction

Organizations increasingly rely upon small groups and teams as the locus of work (Mathieu, Hollenbeck, Van Knippenberg, & Ilgen, 2017). However, we are often surprised by which teams perform well and which do not, since we lack good tools for predicting group performance. By contrast, for over 100 years, psychology has had good tools for predicting many kinds of individual performance, including—especially—individual intelligence tests (Schmidt & Hunter, 1998). Despite controversy about what causes individual intelligence or whether intelligence tests are systematically biased against certain groups, the ability of individual IQ scores to predict an individual's future performance on a wide range of mental tasks remains one of the most replicated results in all of psychology (Deary, 2012).

Motivated by this work, Woolley, Chabris, Pentland, Hashmi and Malone (2010) explored the question of whether something analogous to individual intelligence exists for groups. Is there a similar factor for groups that predicts how well a group will perform a wide range of different tasks? In a series of studies, the researchers found evidence for a single statistical factor for a group that explains approximately 30-50%

Electronic copy available at: https://ssrn.com/abstract=3187373

of the variance in the group's performance on a variety of different tasks. By analogy with individual intelligence (g), they called this factor *c* for "collective intelligence." (Woolley et al., 2010). Credé and Howardson (C&H; 2017) argue that the inference of a general collective intelligence construct is unwarranted because this *c* factor doesn't correlate strongly with all the tasks performed by groups and because of other "statistical artifacts" in the study design and resulting data.

In this paper, we will first clarify how we define collective intelligence (CI) and summarize the accumulated evidence for the validity of the CI construct. Then we will respond specifically to C&H's criticisms of the previous work on collective intelligence. Finally, we will discuss how to further develop the CI construct to strengthen its measurement.

2. What is collective intelligence?

Whether it is valid to say that collective intelligence exists depends, of course, on how we define collective intelligence. There are a number of definitions of CI for which it is obvious that collective intelligence exists, even without any special testing (e.g., see Malone, 2018; Malone & Bernstein, 2015). But in this paper, we focus on the definition of collective intelligence used by Woolley et al (2010):

> *"the general ability of a group to perform a wide variety of tasks."*

This is one way of describing what we might call "general factor intelligence" in which a single general factor predicts the performance of members of a whole population across a wide range of tasks. With IQ tests of individual humans, for example, there is a large group of tasks ("mental tasks") and a population of individuals (humans) for which the individuals' abilities to perform the tasks are predicted by a single statistical factor.

In order to determine whether general factor intelligence exists for a given population and range of tasks, systematic statistical analysis is needed. The first person to do this for individual intelligence was Spearman (1904). Using an early equivalent of what we now call factor analysis, he (and many later researchers) found that (a) a single statistical factor predicted a substantial part of the variance in individual performance on a wide range of mental tasks, (b) no other factor predicted nearly this much, (c) most of the tasks loaded strongly on the first factor, and (d) most of the tasks were positively correlated with one another (Deary, 2000).

It is important to realize that factor analysis does not always lead to a result like this. For example, standard personality variables measuring the characteristics of an individual do not load onto a single general factor. People's tendency to be extraverted, for instance, appears to be uncorrelated with their tendency to be agreeable (Digman, 1990; McCrae & Costa, 1987). And even within the domain of mental tasks, recent research suggests that certain kinds of face recognition tasks are

Electronic copy available at: https://ssrn.com/abstract=3187373

not correlated with the kind of general intelligence measured by IQ tests (Shakeshaft & Plomin, 2015).

In other words, it is an empirical question whether general factor intelligence exists for human groups, not just for individuals, and if so, what is the domain of tasks over which it extends. That is the question that Woolley et al (2010) set out to answer.

2.1. How is the *c* factor related to previous measures used in organizational research on groups and teams?

The kind of collective intelligence measured by Woolley et al (2010) differs in at least two important ways from previous measures used to evaluate teams in organizational research.

First, CI is broader in scope than most existing measures used to evaluate teams. Most self-report surveys, for instance, aim to measure relatively narrowly defined aspects of group functioning with highly overlapping items, and in these cases, psychometricians look for very high average inter-item correlations. Measures of CI, on the other hand, are focused on using a more diverse array of tasks to predict a group's ability to perform a wide range of different tasks together. Thus the overlap among items will be considerably less than one would observe with the more focused measures typically used in teams research. The construct of CI is also much broader than the single-task measures of team performance typically used in laboratory-based research (such as creativity or decision making tasks). At the same time, CI is narrower than more general concepts of team effectiveness, which typically encompass not only team performance but also the well-being of members and the socioemotional processes of the group (Hackman, 1987; Mathieu, Maynard, Rapp, & Gilson, 2008).

Second, in addition to the predictive validity of CI, recent studies have also begun to explore its convergent and discriminant validity vis-à-vis other commonly studied group-based states and processes. Across a series of studies, some published and others under review, researchers have observed a dissociation between CI and indices of group climate or group member relationships. They have found no correlation between CI and group satisfaction (Chikersal, Tomprou, Kim, Woolley, & Dabbish, 2017; Engel, Woolley, Jing, Chabris, & Malone, 2014), relationship quality (Woolley & Aggarwal, under review), or psychological safety (Glikson, Harush, Kim, Woolley, & Erez, 2016; Woolley et al., 2010). By contrast, researchers do see a strong association between CI and transactive memory systems (Kim, Aggarwal, & Woolley, 2016) and some forms of group learning (Aggarwal, Woolley, Chabris, & Malone, 2015; Woolley & Aggarwal, under review). Past research on groups has drawn distinctions between the task vs. socio-emotional processes of a group (Hackman, 1987; Marks, Mathieu, & Zaccaro, 2001; Mesmer-Magnus & DeChurch, 2009), and given the evidence in the literature thus far, it would appear that CI is much more deeply connected to task and cognitive processes in groups than to its socio-emotional processes, further distinguishing CI from more general constructs of group effectiveness (Hackman, 1987; Mathieu et al., 2008)

Electronic copy available at: https://ssrn.com/abstract=3187373

3. Responses to critiques of research on collective intelligence

3.1. Do the relationships among group task scores support the existence of a general factor?

C&H argue that the correlations reported so far among tasks measuring CI are not strong enough to support the existence of a general factor. In particular, they claim that task factor loadings on the CI factor, the average task variance explained by the factor, and the internal consistency among tasks are all too low, and consequently there is no evidence of collective intelligence. To respond to these concerns, we first make a broad conceptual point about the stage of development of current measures of collective intelligence. Then we respond to the detailed statistical arguments C&H make by showing how they are inappropriate in this context.

3.1.1. Stage of development of collective intelligence measurement instruments

The most important limitation of this part of C&H's argument is that they are using criteria that are appropriate for evaluating well-developed psychometric instruments, but they are applying these criteria to the very first investigations of whether the construct in question even exists. In a sense, the measurement of CI is now at a stage similar to where the measurement of individual intelligence was soon after Spearman made his initial observations.

For instance, Woolley et al (2010) designed the initial studies to intentionally sample very *different* kinds of tasks, and it was not known at the time those studies were designed which, if any, of the tasks would be associated with a single collective intelligence factor. In spite of this exploratory nature of the initial studies, a factor analysis of all the groups' scores showed that the first factor accounted for 43% of the variance in performance across all the different tasks. This is consistent with the 30–50% of variance typically explained by the first factor in many well-developed batteries of individual cognitive tasks (Chabris, 2007).

C&H are correct to note that it is possible to have a first factor that explains a substantial amount of variance in a number of variables, even when some of those variables are essentially uncorrelated with the factor. But they are incorrect to assume that when some of the variables in a study don't correlate with the first factor that means no general factor exists. It may simply mean that some of the variables included in the study are not in the domain of variables to which the general factor applies. For instance, in the hypothetical data C&H created for their Table 2, the first four tasks do appear to have a common general factor even though the last four variables are not included in the domain of tasks this factor predicts.

This is often the case in early studies of a new construct, and it would be erroneous to interpret such relationships as evidence that a factor (or the construct it measures) does not exist. For example, if Spearman had included certain kinds of face recognition tasks in his original studies, he would probably have found very low correlations between those tasks and the tasks that loaded heavily on *g* (Shakeshaft & Plomin, 2015). If that had occurred, and he had concluded that *g* did not exist as a result, it would have been an enormous loss to the field of psychology.

It should also be noted that over three decades passed between Spearman's initial observations supporting the existence of the intelligence factor and the publication of the first individual intelligence tests that served as accepted and enduring measures of the construct. Modern intelligence tests are the result of additional decades of psychometric refinements that, among other things, involved developing a large repository of test items and selecting those that maximize the validity and reliability of the test.

Despite its relatively nascent stage of development, the initial findings supporting the existence of CI have subsequently been replicated in a number of studies, both published and under review, in both laboratory and field settings, where researchers find consistent evidence of a strong first factor accounting for 30%-50% of the variance in the data (Chikersal et al., 2017; Engel et al., 2015, 2014; Kim et al., 2017; Woolley & Aggarwal, under review; see Woolley, Riedl, Kim, & Malone, 2017 for a meta-analysis).

In addition, researchers also find evidence of the *predictive validity* of this factor for predicting performance on other kinds of tasks not included in the original estimation of the factor. Establishing the predictive validity of a construct is at least as important as examining internal consistency and reliability. For example, CI has been shown to predict future performance on more complex tasks such as those performed in software programming teams (Engel et al., 2015), in student course projects (Engel et al., 2015; Glikson et al., 2016; Kim et al., 2016), and in online video game teams (Kim et al., 2017) months later.

Together the accumulated findings of all these studies, demonstrating reasonable reliability of the developing CI measure as well as its predictive validity, affirm the initial conclusions that the construct is one worth further investigation. Over time, as more CI tasks are developed and tested, researchers will develop a much better understanding of exactly what kind of tasks are included in the domain of tasks for which a collective intelligence factor exists.

3.1.2. Statistical concerns with C&H's arguments

Even if C&H were evaluating results from measurement instruments that had already undergone many generations of testing and refinement, rather than the early stage exploration of Woolley et al. (2010), the statistical criteria they apply to the collective intelligence data are inappropriate.

First, C&H's sole reliance on shared variance to evaluate validity is problematic. It is well-established in psychometric theory that solely maximizing the shared variance among items almost invariably produces a scale that is too narrow in content, and thus undermines construct validity (i.e., the classic attenuation paradox in psychometrics; Loevinger, 1954, 1957). Therefore, researchers should not construct or evaluate measures solely based on factor analysis or the examination of correlation matrices (Bollen & Lennox, 1991; Briggs & Cheek, 1986; Epstein, 1983). Given the breadth of the CI construct, the researchers intentionally sampled a variety of tasks, in order to cover various aspects of CI. Moreover, despite this task heterogeneity, as described above, they found that CI predicts various criteria, such as group future performance. Thus, C&H's argument, based only on shared variance, is too narrow and does not nullify the predictive validity of CI.

Second, C&H's criteria for the sizes of correlations between tasks and for task factor loadings are much too high and thus inappropriate for a broad construct like CI. C&H argue that a .20 average correlation is too weak to conclude that tasks measure the same latent construct. However, in the literature on psychometrics and task development, it is well-established that moderate inter-item correlations are better than very strong correlations, and that the optimal sizes of inter-item correlations depend on the breadth of the target construct. Typically, inter-item correlations between .20 and .40 are considered optimal (e.g., Bollen & Lennox, 1991; Briggs & Cheek, 1986; Cattell, 1965; Epstein, 1983), as explained by Epstein (1983):

> *What is often not realized is that the average inter-item correlation for most intelligence tests is between .20 and .30. If items in a scale were more highly correlated with each other, they would be too redundant to sample efficiently the breadth of broad personality variables such as intelligence, honesty, or extraversion. An ideal item in a test that measures a broad trait is one that has a relatively high correlation with the sum of all items in the test (minus itself) and a relatively low average correlation with the other items (p. 366).*

Third, C&H also argue that task factor loadings should be greater than .70 (i.e., at least 50% of the variance explained by the general factor). However, in the literature on psychometrics and task development, factor loadings greater than .40 are considered acceptable (Clark & Watson, 1995; Cliff & Hamburger, 1967; Osborne & Costello, 2009; Santor et al., 2011).

We also note that papers that C&H cited as satisfying their criteria used a hierarchical factor model. In those papers, several latent factors were introduced and used to derive a second-order general factor (in contrast to the CI studies, which only modeled a single general latent factor). Therefore, the papers containing these hierarchical factor models that C&H cited are not comparable with the current CI studies. A more reasonable and appropriate source from which to derive criteria thresholds would be the literature on general intelligence. As already discussed, the construct of general intelligence is itself based on the observation that the first factor derived from a battery of cognitive tasks typically explains 30-50% of the variance (Chabris, 2007). Given that the first factor derived from the tasks used to measure CI explain a similar level of variance (i.e., 30-50%), it would seem there is initial evidence for a general factor of collective intelligence.

Finally, we note that C&H's estimates of the shared variance are biased downward by their use of data from sample 6, a conference paper by Barlow & Dennis (2014), which has several problems. First of all, the study only used 3 tasks to measure CI (which is inherently less reliable than the larger sets of tasks used in the other samples). However, C&H included another task, the criterion task in that study (that was administered separately and differently from the other 3 group tasks), in the measure of CI. As a result, the factor model that C&H calculated yields weaker estimates for sample 6. This is important, since most of the low correlations ($r \leq 0$) that C&H report in Figure 1 are derived from sample 6. When dropping sample 6, the average correlation among tasks in published articles is .24, and well within recommended ranges for broad scope measures discussed above.

3.2. Is collective intelligence a statistical artifact caused (or inflated) by low effort responding?

C&H argue that some groups in the studies they reviewed might be exceedingly low in motivation, which could lead to inflated correlations among CI tasks. In principle, they are correct that if (a) some groups have generally low motivation, and (b) low motivation decreases a group's scores on most tasks, then (c) this cluster of consistently low scoring groups together with other groups with uncorrelated task scores would lead to a falsely inflated correlation among the variables where no relationship exists. However, there is strong evidence this hypothesis does not explain what is happening in the studies of CI.

To examine this argument, we conducted a simulation to generate a dataset that conforms to the characteristics that C&H assert are true of the studies of CI. Two clusters of data were generated and for each group, two variables were generated to have almost zero correlation. For the majority group (N=500), each of two variables (i.e., task 1 and 2) was generated from normal distribution with mean of 50 and standard deviance of 15. The minority group (N=50), each of two variables was generated to follow normal distribution with mean of 5 and standard deviation of 1.5.

First, and most important, if the pairwise correlations among task scores are inflated by very low effort by some groups, we would expect to see a pattern similar to that from the simulated data shown in Figure 1a; a cluster of scores in the bottom left corner and a second cloud of uncorrelated points in the middle. This indicates that some teams have exceedingly low scores on both tasks, thus inflating the correlations. But, both in our original data and now with the accumulated data of 758 teams, we see no evidence of this (see Figure 1b). Thus the empirical data does not support the conclusion that the observed correlations are inflated by a cluster of low effort groups.

Second, there is additional empirical evidence that low motivation of group members did not cause low group CI scores. In Study 2 of Woolley et al, 2010, group members completed a validated measure of motivation (Wageman, Hackman, & Lehman, 2005) to indicate how strongly they were motivated for their team to perform well. As reported in the Supplementary Online Materials for that study, the correlation between

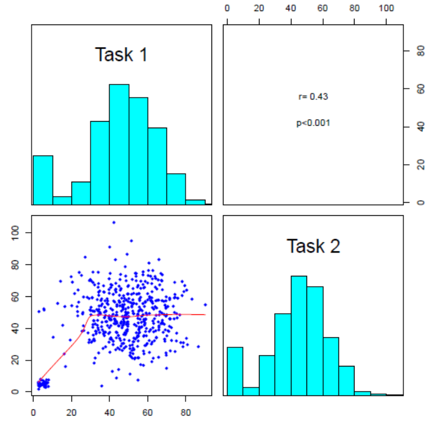

Figure 1a. Simulated data that fit the pattern argued for by Crede and Howardson (2017)

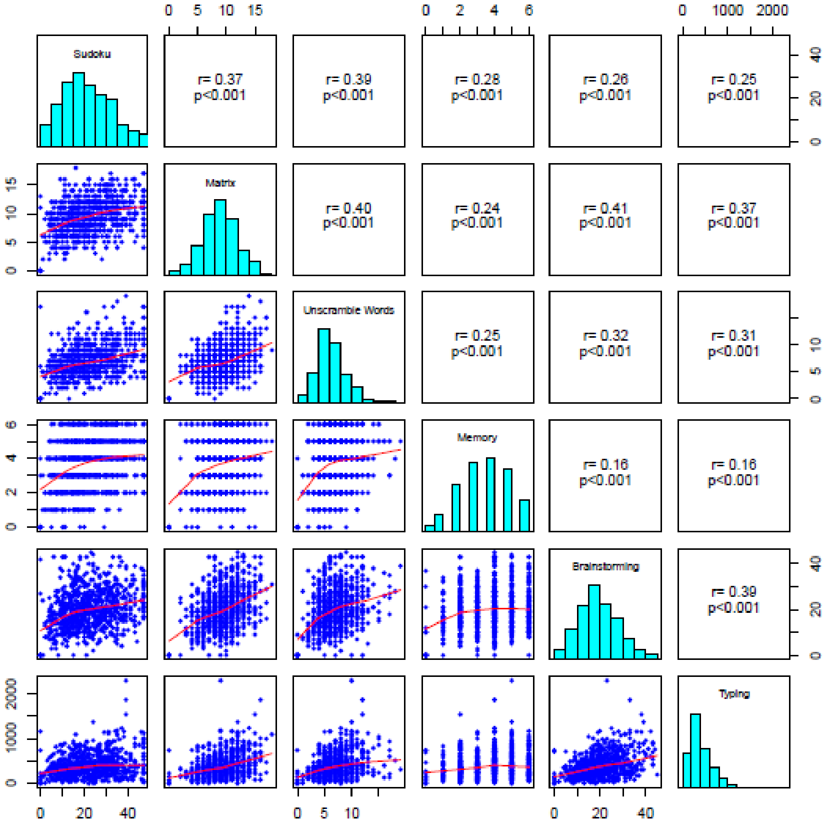

Figure 1b. Pairwise scatterplots from 758 teams in our studies.

the average of this measure for all group members and the group's collective intelligence score was -0.01, *ns*. If C&H's theory about low effort responding had been correct, this correlation should have been strongly, and significantly, positive.

Third, C&H assume that very low scores result from lack of effort. However, they overlook key details in the scoring in Woolley et al. (2010) and subsequent studies that provide a more plausible explanation for these scores. For example, points are deducted when groups performing the typing task skip words or when groups performing the brainstorming task repeat ideas that have already been given. In these cases, a score of zero can be evidence of poor coordination, not, as C&H assumed, lack of effort.

Finally, as another form of evidence that some group members had very low motivation, C&H noted the instance of low Wonderlic Personnel Test (WPT) scores as evidence of low effort, and calculated expected variance of WPT using the central limit theorem (CLT). We realized in reviewing their argument that one of the tables in the supplement to Woolley et al. (2010) had a misleading label, and the range of scores reported were for individuals in the sample, not teams. The WPT scores for individuals in the Woolley et al. (2010) sample have a median of 24, and standard deviation of 6.52. These scores fall between the comparable scores in the norming samples reported by Wonderlic (2002) for the adult working population (median = 22, SD = 7.6) and college graduates (median = 30, SD = 6.3), and these are the two populations from which the Woolley et al. (2010) sample was drawn. Thus, there is nothing to suggest that the Woolley et al. (2010) sample is not representative of the population, or exhibiting exceedingly low effort.

Furthermore, in evaluating the distribution of WPT scores, C&H made assertions about the non-random assignment of participants to teams using calculations based on CLT. However, their calculations are problematic. The CLT states that given *a sufficiently large sample* drawn from a population (e.g., $n > 30$), the mean of *independent and identically distributed (iid) samples* from the same population will approximate the mean of the population. The sample size of $n = 4$, which C&H used in their calculation, is not sufficiently large, and the samples in this context cannot be approximated by iid considering the sampling without replacement in a small, finite sample (Casella & Berger, 2002).

3.3 Is collective intelligence a statistical artifact caused (or inflated) by the influence of nested data?

C&H further argue that the conclusions of a CI factor are inflated by nested data. To demonstrate their argument, they simulated individual-level data and conducted multi-level factor analysis. C&H's simulations demonstrate a well-known phenomenon, the ecological fallacy (Hox, Moerbeek, & van de Schoot, 2010), which occurs when individuals are nested in groups, and when we cannot correctly infer relationships among variables at the group level because the outcome of interest is measured fundamentally at the individual-level (e.g., achievement scores of students, who are nested within school). These issues are a common problem in organizational research, which is inherently multi-level (Rousseau, 1985).

However, C&H's simulations do not show that this phenomenon explains the collective intelligence results in the studies upon which they are commenting. Instead, C&H merely *assume* that there are no group level influences in performance and then simulate what happens under that assumption. More specifically, they assume that the groups determine their answers by having each group member perform the tasks independently and then "averaging" their answers to produce the group answer. However, there are two major flaws with the reasoning underlying this argument.

First, this is far from an accurate model of how groups complete most of the tasks used in the studies they are modeling. Of the full set of ten tasks used in study 2 of Woolley et al. (2010), at least half cannot be readily disaggregated into individual components that could be merely assembled without group interaction, as in a nominal group paradigm (Thorndike, 1938). As described above, for instance, the group typing task requires group members to not only type different sections of the text passage, but also to coordinate their work so no one types the same segment of the text twice and so that there are no gaps between the segments typed. The moral reasoning task requires that groups come to a consensus that balances the interests of different parties to the conflict. And in the brainstorming task, members must not only generate possible answers, but also coordinate their work to avoid duplicates. Under such conditions, some groups will do well less than the sum of their parts because of process loss, due to the failure to coordinate well (Steiner, 1972) while others might exhibit significant synergistic gains (Larson, 2010) and outperform even their best members.

In other words, if groups could in fact disaggregate tasks into individual components, there are some groups (we would argue those that are low in CI) that would perform better if they completed the tasks individually, since working together reduced their aggregate ability! Meanwhile, high CI groups have been shown to consistently operate above the capability of even their best members (Woolley & Aggarwal, under review). However, because some groups gain from interaction while others lose, statistically the numbers could look similar to what C&H produced under their assumption of no interaction, but this does not capture the underlying phenomenon.

A second problem with the assumptions underlying C&H's simulated data is that if the results of Woolley et al. (2010) and the other studies reviewed are merely an aggregate of individual inputs, then features of group interaction should not matter at all to the quality of what groups produce. However, some of the strongest associations observed in Woolley et al. (2010) and subsequent studies relate to the relationship between group communication patterns and CI, as well as individual characteristics that affect the quality of interpersonal interaction such as social perceptiveness. Multiple studies find that the amount and distribution of communication are strong predictors of collective intelligence, even in groups collaborating online via text chat (Engel et al., 2014; Woolley et al., 2010). Social perceptiveness, a characteristic describing how effective individuals are at interpreting subtle cues from interaction partners also ends up being a strong predictor of CI whether groups are working face-to-face or online (Chikersal et al., 2017; Engel et al., 2014; Kim et al., 2017; Woolley, Riedl, et al., 2017). The consistency of these findings suggests that team interaction does play an important role in shaping CI, and thus that C&H's hypotheses to the contrary are not supported by the data

C&H recommend that future research on collective intelligence should measure both the individual-level and group-level performance on the *same* tasks, so that researchers can apply multi-level models. That is an interesting idea, and worth considering; it would enable researchers to estimate what portion of CI is due to individual ability vs. group coordination, a common approach to conceptualizing intelligence in organizations and other larger systems in other disciplines (Knott, 2008). This could enable exploration of the degree to which CI enables synergistic gains in groups (Curşeu, Meslec, Pluut, & Lucas, 2015; Larson, 2007) which research in progress suggests is a good possibility (Woolley & Aggarwal, under review).

2 Conclusions and Moving Forward

We believe we have shown that the arguments C&H made do not, in fact, refute the conclusion that a collective intelligence factor exists. However, the process of devising measurements of the CI factor is still in an early stage, and we completely agree that there is much additional work to be done.

For example, more research is clearly needed on the development and strengthening of measures of CI. In doing so, it will be important to maximize the content validity of measurement by sampling different group task types, while balancing the competing need to maintain an adequate level of internal consistency and reliability. The work to date has focused on sampling from the task domains identified by teams researchers in the past (Larson, 2010; McGrath, 1984; Steiner, 1966). Thus far, studies have been measuring a single factor. As the work evolves, however, it is likely that a two-level factor structure like that often used in individual intelligence (e.g., Deary, 2012) will become useful. In the case of groups, the lower level factors might separately predict performance on several of the major task types identified in existing team task taxonomies, in essence identifying some specialized intelligences in groups.

As researchers continue to develop tasks, they can also more systematically investigate how many tasks should be used to evaluate each task type. The Platform for Online Groups Studies (POGS; Engel et al., 2015; Kim et al., 2017) that has been made available as open source software to other researchers (pogs.mit.edu) facilitates the development and testing of new group-based tasks in different environments. We hope that this instrument can be used and refined by many other researchers.

Another important aspect of continuing to develop our understanding of the CI construct is to evaluate how stable CI is in teams, which is both an important theoretical question as well as a measurement question. There is preliminary evidence, not yet published (Woolley, Kim, Kim, & Malone, 2017) to suggest that CI is relatively stable. In ongoing studies, the observed test-retest reliability of the CI is .88 in teams of strangers completing two sets of tasks one hour apart, .75 in teams of strangers completing them 2 weeks apart (with no interaction in between), and .73 in student teams from academic courses taking them 6-8 weeks apart (with frequent interaction in between). By comparison, some estimates of the test-retest reliability for commonly used measures of individual IQ range between .57 and .94 (Goodman, Streiner, & Woodward, 1974; Snow, Tierney, Zorzitto, Fisher, & Reid, 1989).

In summary, the work on collective intelligence is still at an early stage but shows promising signs of providing a vehicle for the field to conceptualize and measure the collective ability of groups or even larger entities to work together. We are hopeful that we and many other researchers can continue to accumulate data and conduct analyses which will further develop our understanding of the construct, its measurement, and its utility in advancing the science of groups.

3 References

Aggarwal, I., Woolley, A. W., Chabris, C. F., & Malone, T. W. (2015). Cognitive diversity, collective intelligence, and learning. In *Proceedings of Collective Intelligence 2015*. Santa Clara, CA.

Barlow, J. B., & Dennis, A. R. (2014). Not as smart as we think: A study of collective intelligence in virtual groups. In *Proceedings of Collective Intelligence 2014*. Cambridge, MA.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106–148.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.

Cattell, R. B. (1965). *The scientific study of personality*. New York, NY: Penguin Books.

Chabris, C. F. (2007). Cognitive and neurobiological mechanisms of the law of general intelligence. In M. J. Roberts (Ed.) (pp. 449–491). Hove, UK: Psychology Press.

Chikersal, P., Tomprou, M., Kim, Y. J., Woolley, A., & Dabbish, L. (2017). Deep structures of collaboration: Physiological correlates of collective intelligence and group satisfaction. *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309.

Cliff, N., & Hamburger, C. D. (1967). The study of sampling errors in factor analysis by means of artificial experiments. *Psychological Bulletin*, *68*(6), 430.

Credé, M., & Howardson, G. (2017). The structure of group task performance—A second look at “collective intelligence”: Comment on Woolley et al. (2010)., *102*(10), 1483–1492.

Curşeu, P. L., Meslec, N., Pluut, H., & Lucas, G. J. M. (2015). Cognitive synergy in groups and group-to-individual transfer of decision-making competencies. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01375

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain*. New York: Oxford University Press.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*(1), 453–482.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417–440.

Engel, D., Woolley, A. W., Aggarwal, I., Chabris, C. F., Takahashi, M., Nemoto, K., … Malone, T. W. (2015). Collective intelligence in online collaboration emerges in different contexts and cultures. In *CHI '15 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Seoul, Korea.

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. https://doi.org/10.1371/journal.pone.0115212

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360–392.

Glikson, E., Harush, R., Kim, Y. J., Woolley, A. W., & Erez, M. (2016). *Psychological safety and collective intelligence in multicultural globally dispersed teams*. Paper presented at the Interdisciplinary Network for Groups Research (INGRoup) conference, Helsinki, Finland.

Goodman, J. T., Streiner, D. L., & Woodward, C. A. (1974). Test-retest reliability of the Shipley-Institute of Living Scale: Practice effects or random variation. *Psychological Reports*, *35*(1), 351–354.

Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315–342). Englewood Cliffs, NJ: Prentice Hall.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Kim, Y. J., Aggarwal, I., & Woolley, A. W. (2016). How communication impacts team performance: Exploring collective intelligence and transactive memory system as mechanisms. Presented at the Annual Convention of the International Communication Association, Fukuoka, Japan.

Kim, Y. J., Engel, D., Woolley, A., Lin, J., McArthur, N., & Malone, T. (2017). What makes a strong team? Using collective intelligence to predict performance of teams in League of Legends. *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*.

Knott, A. M. (2008). R&D/Returns causality: Absorptive capacity or organizational IQ. *Management Science*, *54*(12), 2054.

Larson, J. R. (2007). Deep diversity and strong synergy: Modeling the impact of variability in members' problem-solving strategies on group problem-solving performance. *Small Group Research*, *38*(3), 413–436. https://doi.org/10.1177/1046496407301972

Larson, J. R. (2010). *In Search of Synergy in Small Group Performance*. New York, NY: Psychology Press.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694.

Malone, T. W. (2018). *Superminds: The suprising power of people and computers thinking together*. Boston: Little, Brown and Company.

Malone, T. W., & Bernstein, M. S. (2015). Introduction. In T. W. Malone & M. S. Bernstein (Eds.), *Collective Intelligence Handbook*. Cambridge, MA: MIT Press.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376.

Mathieu, J. E., Hollenbeck, J. R., Van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 452–467.

Mathieu, J. E., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management, 34*(3), 410–476.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90.

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

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology, 94*(2), 535–546.

Osborne, J. W., & Costello, A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review*, *12*(2), 131–146.

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. In L. L. Cummings & B. Staw (Eds.) (Vol. 7, pp. 1–37). Greenwich, CT: JAI.

Santor, D. A., Haggerty, J. L., Lévesque, J.-F., Burge, F., Beaulieu, M.-D., Gass, D., & Pineault, R. (2011). An overview of confirmatory factor analysis and item response analysis applied to instruments to evaluate primary healthcare. *Healthcare Policy*, *7*(Special Issue).

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262.

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12887–12892.

Snow, W. G., Tierney, M. C., Zorzitto, M. L., Fisher, R. H., & Reid, D. W. (1989). WAIS-R Test-retest reliability in a normal elderly sample. *Journal of Clinical and Experimental Neuropsychology, 11*(4), 423–428.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, *11*(4), 273–283.

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Thorndike, R. L. (1938). On what type of task will a group do well? *The Journal of Abnormal and Social Psychology*, *33*(3), 409.

Wageman, R., Hackman, J. R., & Lehman, E. (2005). Team Diagnostic Survey: Development of an instrument. *Journal of Applied Behavioral Science*, *41*(4), 373–398. https://doi.org/10.1177/0021886305281984

Woolley, A. W., & Aggarwal, I. (under review). *Collective intelligence, interpersonal relationships, and group learning*.

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., Kim, Y. J., Kim, Y., & Malone, T. W. (2017). *[Test-retest reliability of the Team Collective Intelligence Test]*. Unpublished raw data.

Woolley, A. W., Riedl, C., Kim, Y. J., & Malone, T. W. (2017). More evidence for a general collective intelligence factor in human groups: A meta-analysis. *Proceedings of Collective Intelligence 2017*.