Metacognitive Confidence and Communication in Dyadic Decision-Making

Matthew David Blanchard

A thesis submitted in fulfilment of the requirements for the degree Doctor of Philosophy

School of Psychology, Faculty of Science

The University of Sydney

2025

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Matthew Blanchard

_____

Date: 23/06/2025

Artificial Intelligence Statement

Generative Artificial Intelligence was used during the preparation of this thesis to assist with copyediting tasks, including grammar refinement, phrasing, and improving clarity. All content was originally written by Matthew David Blanchard, and no material was included without substantial modification and critical oversight by the author. All intellectual contributions, interpretations, and arguments are the author's own.

Acknowledgements

Above all else, I would like to express my infinite gratitude to my primary supervisor, Sabina Kleitman. I would not be here without you. Your wisdom, humour, patience, and unwavering encouragement sustained me through the most challenging stretches of this journey. Whenever I felt lost in ideas or overwhelmed by doubt, you always knew how to gently untangle the mess and guide me back on course. Most of all, I will miss your elaborate April Fool's Day pranks. I wish I could say I won't fall for another, but we both know that isn't true.

Deep gratitude goes to my auxiliary supervisors. Eugene Aidman, thank you for always asking "so what?". A question that pushed me to think beyond my narrow focus and to connect my research to real-world contexts. Lazar Stankov, thank you for your guidance on research design, measurement, and analytical approaches, and for your consistently prompt and constructive feedback. Both your support enhanced the novelty and rigour of the studies conducted in fulfilment of this thesis.

To the colleagues, mentors, and volunteers who made meaningful contributions to this work: thank you. Simon Jackson, who supervised my honours year, helped spark a passion for coding and statistics that burns brighter today. Marvin Law, I will always look back fondly on our time sharing the CODES lab. Those were good days, full of collaboration, constructive and heated debates, laughter, and the occasional bouldering session. I'd also like to thank the volunteers who contributed to the communication coding in Study 1: Benjamin Kai Ni, Fennella Palanca, Daniel Sanchez, and Vivian Nguyen. You generously devoted countless hours to a tedious but essential task. My thanks also go to Ralf Kurvers and the team at the Center for Adaptive Rationality at the Max Planck Institute for Human Development for their valuable feedback on the design of Study 3. And I'm grateful to the

Australian Government's Department of Education and the Defence Science and Technology Group for awarding scholarships that made it possible to begin this journey in my 30s.

Finally, thank you to Adele, Mum, Rhye, Dave, Santi, and Jodie for your enduring support and encouragement throughout the years.

I am deeply indebted to you all.

Author Attribution Statement

Publications Presented in This Thesis

This thesis is composed of three empirical studies that have been accepted or submitted for publication in peer-reviewed journals. Chapter 2 has been published in *Frontiers in Psychology* and Chapter 3 and Chapter 4 have both been submitted to *Cognitive Research: Principles and Implications* (APA journal). The co-authors on these manuscripts are members of my supervisory team: Professor Sabina Kleitman, Professor Eugene Aidman, and Professor Lazar Stankov. The thesis author (M.D.B.) is the leading author for all three manuscripts. The chapter publications and their respective author contributions are listed below.

Chapter 2: Blanchard, M. D., Kleitman, S., & Aidman, E. (2023). Are two naïve and distributed heads better than one? Factors influencing the performance of teams in a challenging real-time task. *Frontiers in Psychology*, *14*, 1042710. https://doi.org/10.3389/fpsyg.2023.1042710.

Author contributions: Conceptualization, M.D.B., S.K., EA, and Simon Jackson; methodology, M.D.B., S.K., EA, and Simon Jackson; formal analysis, M.D.B. and S.K.; investigation, M.D.B.; data curation, M.D.B.; writing the original draft, M.D.B.; writing review and editing, M.D.B., E.A., and S.K.; visualization, M.D.B.; main supervision, S.K.; associate supervision E.A. and L.S.

Chapter 3: Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2024). A Recipe for Dyadic Collective Intelligence for Well-Structured Tasks: Mix Equal Parts Cognitive Ability and Confidence Plus a Pinch of Social Sensitivity. *Cognitive Research: Principles and Implications*, *10*, 63. https://doi.org/10.1186/s41235-025-00655-0.

Author contributions: Conceptualization, M.D.B., S.K., E.A., and L.S.; methodology, M.D.B. and S.K.; formal analysis, M.D.B. and S.K.; investigation, M.D.B.; data

curation, M.D.B.; writing the original draft, M.D.B.; writing review and editing, M.D.B., E.A., L.S., and S.K.; visualization, M.D.B.; main supervision, S.K.; associate supervision E.A. and L.S.

Chapter 4: Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2024). How Trait Confidence and Communication Shape Dyadic Decision Outcomes and Confidence Matching. *Cognitive Research: Principles and Implications.* Manuscript submitted for publication.

Author contributions: Conceptualization, M.D.B., S.K., E.A., and L.S.; methodology, M.D.B. and S.K.; formal analysis, M.D.B.; investigation, M.D.B.; data curation, M.D.B.; writing the original draft, M.D.B.; writing review and editing, M.D.B., E.A., L.S., and S.K.; visualization, M.D.B.; main supervision, S.K.; associate supervision E.A. and L.S.

I confirm that the author attribution statements above are correct.

_____

Matthew David Blanchard

Date: 24/06/2025

As a supervisor for the candidature upon which this thesis is based, I can confirm that the author attribution statements above are correct.

_____

Sabina Kleitman

Date: 24/06/2025

Eugene Aidman	Lazar Stankov
Date: 24/06/2025	Date: 24/06/2025

Additional Journal Articles Published During this PhD Candidature

Stapleton, P., Douglas, A., & Blanchard, M. D. (2025). Daily Meditation Versus Emotional Freedom Techniques: A Pilot Australian Primary School Trial. *International Online Journal of Primary Education*. Accepted for publication.

Blanchard, M. D., Herzog, S. M., Kämmer, J. E., Zóller, N., Kostopoulou, O., & Kurvers, R. H. J. M. (2024). Collective Intelligence Increases Diagnostic Accuracy in a General Practice Setting. *Medical Decision Making*, 44 (4): 451-462. https://doi.org/10.1177/0272989X241241001

Stapleton, P., Wilson, C., Uechtritz, N., Stewart, M., McCosker, M., O'Keefe, T., & Blanchard, M. D. (2024). A randomized clinical trial of emotional freedom techniques for chronic pain: Live versus self-paced delivery with 6-month follow-up. *European Journal of Pain, 29(3).* https://doi.org/10.1002/ejp.4740

Kleitman, S., Fullerton, D. J., Law, M. K., Blanchard, M. D., Campbell, R., Tait, M. A., ... & King, M. T. (2023). The Psychology of COVID-19 Booster Hesitancy, Acceptance and Resistance in Australia. *Vaccines*, 11(5), 907. https://doi.org/10.3390/vaccines11050907

Kleitman, S., Jackson, S. A., Zhang, L. M., Blanchard, M. D., Rizvandi, N. B., & Aidman, E. (2022). Applying Evidence-Centered Design to Measure Psychological Resilience: The Development and Preliminary Validation of a Novel Simulation-Based

Assessment Methodology. *Frontiers in Psychology*, 12, 717568. https://doi.org/10.3389/fpsyg.2021.717568

Kleitman, S., Fullerton, D. J., Zhang, L. M., Blanchard, M. D., Lee, J., Stankov, L., & Thompson, V. (2021). To comply or not comply? A Latent Profile Analysis of Behaviours and Attitudes During the COVID-19 Pandemic. *PloS one*, 16(7), e0255268. https://doi.org/10.1371/journal.pone.0255268

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2020). Collective Decision Making Reduces Metacognitive Control and Increases Error Rates, Particularly For Overconfident Individuals. *Journal of Behavioral Decision Making*, 33(3), 348-375. https://doi.org/10.1002/bdm.2156

Conference Proceedings and Presentations

Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2024, December). *Decision Making in Dyads: Team Performance Depends on Communication, Cognitive Ability, Confidence, and Personality.* Poster presentation at the Defence Human Sciences Symposium (DHSS) in Melbourne, Australia.

Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2023, December). *Confidence and Intelligence Indicate How Much Two Heads are Better Than One.* Oral presentation at the Australian Conference of Personality and Individual Differences (ACPID) in Noosa, Australia.

Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2023, December). *Confidence and Intelligence Indicate How Much Two Heads are Better Than One.* Poster presentation at the DHSS in Brisbane, Australia

Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2022, December). *When are Two Heads Better Than One?.* Oral presentation at the DHSS in Sydney, Australia

Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2020, December). *Are Two Heads Always Better Than One? Performance Cues and Communication Patterns in Naïve Distributed Teams Performing a Dynamic Task.* Oral presentation at the DHSS in Melbourne, Australia

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2019, October). *The Relationship Between Collective Error Rates and Metacognition for Behavioral Decisions.* Oral presentation at the Center for Adaptive Rationality at the Max Planck Institute for Human Development in Berlin, Germany.

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2018, November). *Team Similarity Profiles: Is Psychological Similarity Associated with Team Performance?.* Oral presentation at the DHSS in Perth, Australia

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2018, October). *Collective Decision-Making Increases Error Rate, Particularly for Overconfident Individuals.* Oral presentation at the Conference on Decision Sciences in Konstanz, Germany.

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2018, October). *The Relationship Between Collective Error Rates and Metacognition for Behavioral Decisions.* Oral presentation to the Faculty of Data and Decision Sciences at Technion University in Haifa, Isreal.

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2018, September). *Collective Decision-Making Increases Error Rate, Particularly for Overconfident Individuals.* Oral presentation at EARLI SIG 16 conference in Zurich, Switzerland.

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2017, November). *Collective Decision-Making Increases Error Rates, Particularly for High-Confidence Individuals.* Poster presentation at the Society for Judgment and Decision Making conference in Vancouver, Canada.

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2017, November). *Collective Decision-Making Increases Error Rates, Particularly for High-Confidence Individuals.* Poster presentation at the DHSS in Adelaide, Australia.

Scholarships and Awards

This research was supported by several scholarships including an Australian Government Research Training Program (RTP) Scholarship.

2018 – 2021	Team Performance Top-Up Scholarship funded by The Defence Science and Technology Group (DSTG) ($30,000)
2019	William and Catherine Mcllrath Scholarship ($2,500)
2017 – 2020	RTP Stipend funded by the Australian Government's Department of Education ($95,000)
2017	Adaptable Cognition Top-up Scholarship funded by The Defence Science and Technology Group ($10,000)
2017	Best Mini-oral Presentation at DHSS ($500)

Symbols and Acronyms

Statistical Parameters	
ACME	Average causal mediation effects
ADE	Average direct mediation effects
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
b	Unstandardised regression coefficient
β	Standardised regression coefficient
CFI	Comparative Fit Index
CI	Confidence interval
df	Degrees of freedom
F	*F*-ratio statistic
GFI	Goodness of Fit Index
h^2	Communalities
κ	Fleiss' Kappa
LogLik	Log likelihood ratio test
η^2	Eta squared effect size
η_p^2	Partial eta squared effect size
p	*p*-value
r	Correlation coefficient
RMSEA	Root Mean Square Error of Approximation
R^2	Variance accounted for by fixed effects
ΔR^2	Change in R^2
SD	Standard deviation
SE	Standard error
t	*t*-test estimate
TLI	Tucker- Lewis Index
χ^2	Chi-square
χ^2 / *df*	Relative chi-square
Wilk's Λ	Wilk's lambda for MANOVA
ω_t	Omega total
Other Acronyms	
ADR	Applying decision rules
CFA	Confirmatory factor analysis
CHC	Cattell-Horn-Carroll model of intelligence
CRT	Cognitive reflection test
EFA	Exploratory factor analysis
LMM	Linear mixed-effects model
LPA	Latent profile analysis
MDMT	Medical decision making test
RAPM	Raven's advanced progressive matrices
GT	Geography test
2HBT1	Two heads are better than one

Glossary of Terms

Asymmetric group: A group whose members are assigned different task roles and are exposed to different task conditions. For example, in a driver and navigator setting, the driver controls a vehicle, must navigate through obstacles, and can only see operational conditions from their vehicle's point-of-view. Whereas the navigator of an uninhabited aerial vehicle operator can freely roam the environment independently of the driver and provides the driver with information from their aerial perspective.

Carroll-Horn-Cattel model of intelligence: An empirically validated framework for understanding the cognitive processes underlying individual performance across a wide variety of tasks (e.g., Carroll, 1993; Horn & Cattell, 1966; McGrew, 2009). The model includes 16 broad abilities, such as: Fluid Reasoning, abstract reasoning that has little dependence on acquired knowledge; Crystallised intelligence, acquired knowledge that is culturally relevant; and Quantitative Knowledge, acquired knowledge about mathematics.

Collective intelligence: A group's general ability to perform across a wide variety of tasks (Woolley et al., 2010), similar to general intelligence for individuals.

Confidence matching: A tendency for group members to align their decision confidence levels over time (Bang et al., 2017).

Confirmatory factor analysis: A statistical technique used to test whether a hypothesised factorial structure fits a set of observed variables (Kline, 2014).

Decision confidence: Confidence judgments made for responses to items within a specific task. Ratings reflect moment-to-moment monitoring of one's performance and guide collective decisions in real-time (e.g., Koriat, 2008).

Distributed group: A group whose members are not physically co-located, with communication mediated by technology.

Dyad: A group with two members.

Dynamic task: A task with operational conditions that change within trials.

Ill-structured task: Tasks that have open ended responses, multiple correct solutions, and multiple pathways to completion (Laughlin, 2011). Also known as judgemental tasks**.**

Latent profile analysis: A mixture modelling statistical technique that classifies individuals into distinct subgroups based on a set of observed variables (Spurk et al., 2020).

Linear mixed-effect model: A multilevel modelling technique that extends linear regression by incorporating both fixed (e.g., experimental conditions) and random effects (e.g., individual differences) to account for the hierarchical structure and non-independence of data (Gelman & Hill, 2007).

Static task: A task with operational conditions that remain constant within trials.

Trait confidence: A stable, domain-general tendency for confidence judgments across different tasks and contexts, relative to others (e.g., Johnson, 1939; Kleitman & Stankov, 2001). Typically derived from multiple decision confidence measures taken across different cognitive domains (Stankov et al., 2014). Thus, trait confidence reflects broader individual differences in metacognitive self-beliefs.

Two heads are better than one: The phenomenon where two people working together typically perform better than the average of the two individuals working alone.

Well-structured task: Tasks with clearly defined goals and a single correct solution for each item (Laughlin, 2011). Also known as intellective tasks.

Outline of Studies

Study	Chapter	Sample	Measures	Procedure
1	2	N = 294 Australian undergraduate psychology students (196 females, 98 males, mean age = 19.80, SD = 4.12). Participants were assigned to either the individual (*n* = 134) or dyadic (*n* = 80) condition. Grouping (individual vs dyad) was a between-subjects factor.	<u>Collective</u> 1. Driving Simulation (5 trials) <u>Individual</u> 2. Raven's Advanced Progressive Matrices (20 items) 3. Random Number-Letter Switching Test (72 items) 4. Flanker test (100 items) 5. Running Letter Span (15 items) 6. Mini International Personality Item Pool (Mini-IPIP; 20 items)	Completed in a university computer lab with up to four participants per two-hour session. Participants were randomly assigned to a condition upon arrival. Tasks were completed in a fixed order, with more cognitively demanding tasks administered earlier to reduce fatigue effects.
2	3	N = 210 Australian undergraduate psychology students (133 females, 77 males, mean age = 20.79, SD = 4.33) completed the study as 105 dyads. Grouping was a repeated-measures factor.	<u>Collective</u> 1. Applying Decision Rules (10 items) 2. Cognitive Reflection Test (7 items) 3. Geography test (11 items) 4. Ravens Advanced Progressive Matrices (18 items) <u>Individual</u> 1. Composite Emotions Task (36 items) 2. Medical Decision Making Test (16 items) 3. Running Letter Span (15 items) 4. Mini-IPIP (20 items) 5. Ravens Advanced Progressive Matrices (18 items)	Completed in a university computer lab with up to four participants per two-hour session. Participants were randomly paired into dyads upon arrival. Task order was counterbalanced to reduce practice and fatigue effects.

3	4	N = 210 Australian psychology students (158 females; Mean age = 22.02, SD = 6.12) who completed the study in 105 dyads. Grouping was a repeated-measures factor.	<u>Collective</u> 1. Three General Knowledge Tests (10 items each) under isolated, passive, and active communication conditions <u>Individual</u> 2. Ravens Advanced Progressive Matrices (36 items) 3. Mini-IPIP (20 items) 4. Esoteric Analogies Test (20 items) 5. Social Motivation Scale (57 items) 6. Trust Scale (5 items) 7. Psychological Safety Scale (7 items) 8. Empathy Quotient (60 items) 9. Reading the Mind in the Eyes (10 items) 10. BIS/BAS (24 items) 11. Risk aversion (10 items)	Completed remotely via Zoom with two participants per two-hour session. Dyads were pre-paired based on trait-confidence (measured in a pre-screening study). Communication conditions were counterbalanced across sessions. The individual measures were presented in a fixed order.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 BACKGROUND 1

1.2 RESEARCH AIMS 4

1.3 GENERAL METHOD 5

1.4 THESIS STRUCTURE 6

CHAPTER 2: STUDY 1 8

2.1 INTRODUCTION 8

2.1.1 Are 2HBT1 Under Different Operational Conditions? 10

2.1.2 Team Communication 12

2.1.2.1 Measurement 13

2.1.2.2 The Cost of Communication 15

2.1.2.3 Team Communication in the Present Study 16

2.1.3 Simulation-Based Assessment 17

2.1.4 The Role of Individual Differences 18

2.1.5 The Present Study 19

2.1.6 Statistical Analyses 22

2.2 METHOD 23

2.2.1 Participants 23

2.2.2 Measures 23

2.2.3 Driving Simulation 26

2.2.4 Quantifying Team Communication 27

2.2.4.1 Categories of Communication Behaviour 27

2.2.4.2 Coding communication 30

2.2.5 Procedure 31

2.3 RESULTS....32
2.3.1 Descriptive Statistics....32
2.3.1.1 Simulation Derived Performance Metrics....32
2.3.1.2 Communication Measures....34
2.3.1.3 Individual Differences Measures....36
2.3.2 Exploratory Factor Analysis using Communication Variables....36
2.3.3 Performance During the Simulation....39
2.3.3.1 Individuals vs. Dyads....39
2.3.3.2 Communication as a Predictor....40
2.3.3.3 Volume of Communication....43
2.4 DISCUSSION....45
2.4.1 Two Heads are Not Always Better Than One....45
2.4.2 Communication and Dyadic Performance....47
2.4.3 Comparing Communication Metrics....48
2.4.4 Covariates and Dyadic Performance....49
2.4.5 Limitations and Future Directions....50
2.4.6 Implications....51
2.4.7 Conclusion....52
CHAPTER 3: STUDY 2....54
3.1 INTRODUCTION....54
3.1.1 Collective Intelligence....55
3.1.1.1 Original Research Findings....56
3.1.1.2 Independent Research and Meta-Analytic Findings....57
3.1.1.3 Methodological Considerations....60
3.1.1.4 The Current Research....62
3.1.2 Confidence and Group Decision-Making....63

3.1.3 Profiling Dyadic Performance *65*

3.1.4 Individual Differences *66*

3.1.5 The Present Study *67*

3.2 METHOD 68

3.2.1 Participants *68*

3.2.2 Measures *69*

3.2.2.1 Collective Intelligence Tasks 69

3.2.2.2 Other Measures 71

3.2.3 Communication Measures *73*

3.2.4 Procedure *73*

3.3 RESULTS 74

3.3.1 Descriptive statistics *74*

3.3.1.1 Accuracy and Confidence 74

3.3.1.2 Individual Difference and Communication Measures 76

3.3.2 Extracting Collective Intelligence and Confidence factors *76*

3.3.3 The Predictors of Collective Intelligence and Confidence *79*

3.3.4 Mediation Analysis for the Proportion of Females, Social Sensitivity, and Collective Intelligence *81*

3.3.5 Latent Profile Analysis *83*

3.3.5.1 Selecting an LPA Solution 83

3.3.5.2 Interpretation of the 3-Class Solution 83

3.3.5.3 Differences Between the Three Profiles 84

3.4 DISCUSSION 87

3.4.1 The Predictors of Collective Intelligence for Well-Structured Tasks *87*

3.4.2 The Psychological Profiles of Dyads *91*

3.4.3 Implications, Limitations, and Future Directions *92*

3.4.4 Conclusion *95*

CHAPTER 4: STUDY 3 ..... **96**

4.1 INTRODUCTION 96

4.1.1 Collective Decisions *96*

4.1.2 The Confidence Theory *98*

4.1.3 Confidence Matching *99*

4.1.4 The Role of Trait Confidence *101*

4.1.5 Profiling Dyadic Decision-Making *104*

4.1.6 The Present Study *105*

4.2 METHOD 110

4.2.1 Main Study *111*

4.2.1.1 Participants 111

4.2.1.2 Measures 111

4.2.1.3 Communication Measures 113

4.2.1.4 Procedure 114

4.2.1.5 Statistical Analyses 115

4.3 RESULTS 116

4.3.1 Descriptive Statistics *117*

4.3.1.1 Decision Accuracy and Decision Confidence 117

4.3.1.2 Demographic and Individual Difference Measures 118

4.3.2 Effects of Trait Confidence and Communication on Decision Outcomes *120*

4.3.2.1 Baseline Analyses 121

4.3.2.2 Main Analyses 122

4.3.2.2.1 Decision Accuracy 122

4.3.2.2.2 Decision Confidence 123

4.3.2.3 Confidence Matching....125
4.3.3 The Psychological Profiles of Dyads....129
4.3.3.1 Selecting an LPA solution.....130
4.3.3.2 Interpretation of the Four Profiles....131
4.3.3.3 Differences Between the Four Profiles....131
4.4 DISCUSSION....135
4.4.1 Decision Accuracy: Interaction of Communication and Trait Confidence 135
4.4.2 Decision Confidence: Passive Gains and Active Uncertainty....137
4.4.3 Confidence Matching: Communication Mode Matters....138
4.4.4 Distinct Psychological Profiles of Dyads....140
4.4.5 Implications and Contributions to Theory....141
4.4.6 Practical Implications....142
4.4.7 Limitations and Future Directions....143
4.4.8 Conclusion....144
CHAPTER 5: GENERAL DISCUSSION....145
5.1 GENERAL OVERVIEW....145
5.2 MAIN CONTRIBUTIONS TO THE EXISTING LITERATURE....146
5.2.1 Cognitive Ability and Effective Collaboration....146
5.2.2 Trait Confidence as a Moderator....147
5.2.3 Communication and its Interaction with Trait Confidence....149
5.2.4 The Psychological Profiles of Dyads....151
5.3 KEY CONTRIBUTIONS....152
5.3.1 Extension of the Confidence Theory....152
5.3.2 Person-Centred Statistical Analyses....153
5.3.3 Use of Different Tasks and Response Options....153

5.4 General Limitations 154

5.4.1 The Unit of Analysis Problem 154

5.4.2 Limited Generalisability Based on Group Size 156

5.4.3 Limited Consideration of Gender Effects 156

5.4.4 Limited Generalisability Based on Task Types 157

5.5 Implications and Future Directions 158

5.5.1 Theoretical Implications 158

5.5.2 Methodological Implications 159

5.5.3 Selection and Training 159

5.6 Conclusion 161

List of Tables

Table 2.1. The Communication Coding System....28

Table 2.2. Descriptive Statistics and Internal Consistency Estimates for Simulation Performance (N=80) 33

Table 2.3. Descriptive Statistics and Internal Consistency Estimates for the Communication Behaviours (N=53 Dyads) 34

Table 2.4. Descriptive Statistics and Internal Consistency Estimates for Volume-Based Communication (N=53 Dyads)....36

Table 2.5. Communication Intercorrelations and EFA Results for the Normal Condition....37

Table 2.6. Communication Intercorrelations and EFA Results for the Fog Condition....38

Table 2.7. Results of ANOVAs for Collisions and Speed 40

Table 2.8. Hierarchical Regression Analyses Predicting Performance Using the Quality of Communication (N=53 Dyads)....42

Table 2.9. Hierarchical Regression Analyses Predicting Performance Using Volume of Communication Measures (N=53 Dyads) 44

Table 3.1. The Hypotheses for Study 2 68

Table 3.2. Descriptive Statistics and Internal Consistency Estimates for Measures of Accuracy and Confidence (N=105) 75

Table 3.3. Summary of Fit Indices for Different Models of Collective Intelligence and Collective Confidence Using CFA (N = 105) 78

Table 3.4. Summary of Standardised Regression Weights, Communalities, and Correlations for a CFA (N = 105) 79

Table 3.5. Hierarchical Regression Analyses Predicting Collective Intelligence and Collective Confidence (N = 105)....80

Table 3.6. The Results of a Mediation Analysis for Social Sensitivity Mediating the Relationship Between Proportion of Females and Collective Intelligence (N = 105)....82

Table 3.7. Results of ANOVAs for the Differences Between LPA Profiles (N = 105) 86

Table 4.1. Descriptive Statistics and Internal Consistency Estimates for Decision Accuracy and Decision Confidence for Each Condition (N = 210; n = 70 per Trait Confidence Condition) 117

Table 4.2. Descriptive Statistics and Internal Consistency Estimates for Demographic, Individual Difference, and Communication Variables for the Trait Confidence Conditions (N = 210; n = 70 Per Trait Confidence Condition) 119

Table 4.3. Pairwise Comparisons for Decision Accuracy and Confidence (N = 1256) 121

Table 4.4. Main Effects and Simple Interaction Contrasts for Confidence Matching (N = 6280) 126

Table 4.5. Main Effects and Simple Interaction Contrasts for Confidence Matching Predicting the Change in Decision Accuracy (N = 628).....129

Table 4.6. Frequency Table for Membership in Trait Confidence Conditions and the LPA Profiles 132

Table 4.7. Results of ANOVAs for the Differences Between LPA Profiles (N = 105) 134

List of Figures

Figure 1.1. Order of Presentation for Study 2 and Study 3 Items....6

Figure 2.1. Common Communication Metrics....13

Figure 2.2. The Driver's and Navigator's Screens During the Normal Condition (A and C) and Fog Condition (B and D)....27

Figure 2.3. Process of Coding Each Speaking Turn During the Driving Simulation....31

Figure 2.4. The Communication Coding Application....32

Figure 2.5. Frequency Distributions for Individual Collisions (A) and Speed (B) Overall and Dyad Collisions (C) and Speed (D) Overall....33

Figure 2.6. Mean Collisions (A) and Speed (B) for Individuals and Dyads During Both Conditions....40

Figure 2.7. Scatterplots for Harmful Navigator and Collisions (A and B) and Helpful Exchange and Speed (C and D) During Both Conditions....43

Figure 3.1. Hypothesized One First-Order Factor Model (A) and Two First-Order Factors Model (B) Without Modification. Solid Lines Represent Positive Loadings/Correlations....77

Figure 3.2. Mean Scores for the Three Latent Profiles....84

Figure 3.3. Differences Between the Three Latent Profiles....85

Figure 4.1. General Knowledge Test Procedure for Each Communication Condition....115

Figure 4.2. The Differences Between Dyads and Individuals on Decision Accuracy and Decision Confidence for All Conditions....120

Figure 4.3. Confidence Matching for the Communication and Trait Confidence Conditions....127

Figure 4.4. The Relationship Between Confidence Matching and the Change in Decision Accuracy for the (A) Communication and (B) Trait Confidence Conditions....129

Figure 4.5. Mean Scores for the Four Latent Profiles....130

Figure 4.6. Differences Between the Four Latent Profiles....133

Abstract

This thesis investigated the psychological traits and communicative factors that shape collaboration in dyadic decision-making. Across three empirical studies involving 714 participants, it examined the relationships between metacognitive confidence, communication, and dyadic outcomes across dynamic and static task contexts.

Study 1 explored the "two heads are better than one" effect in a dynamic driving simulation with changing operational conditions. Using a novel, behaviour-based coding system, the study found that communication quality, but not quantity, predicted dyadic accuracy, and that dyads outperformed individuals but only under specific conditions.

Study 2 examined the emergence of collective intelligence in dyads completing static tasks guided by the Cattell-Horn-Carroll model of individual intelligence. The results challenged prior claims that collective intelligence is primarily shaped by equal participation in discussions, gender composition, and social sensitivity, and instead highlighted the central roles of individual cognitive ability and metacognitive confidence. Latent profile analysis identified distinct dyad types, each characterised by unique psychological profiles and performance patterns.

Study 3 extended the confidence theory, which proposes that groups share and use expressions of confidence to guide their decisions, by examining how individual differences in trait-confidence and communication modality interact to shape dyadic processes and outcomes. Results showed that dyads with low-trait confidence improved equally under passive (visual information sharing only) and active communication (verbal discussion), whereas those with mixed-trait or high-trait confidence benefited most from active communication. Additionally, dyad members' confidence ratings became more aligned (i.e., confidence matching) under both passive and active communication, but this alignment predicted accuracy improvements only in the passive communication condition.

Together, these three studies demonstrate that no single factor guarantees collaborative success. Instead, dyadic performance is shaped by a dynamic interplay between individual differences in trait confidence, communication type and context, and task structure. This thesis advances theoretical understanding of metacognitive confidence in dyadic decision-making, introduces novel methodologies for capturing communication quality and profiling dyads, and highlights practical applications of the findings.

Chapter 1: Introduction

1.1 Background

Collective decision-making refers to the process by which groups work together to pool information, deliberate, and ultimately arrive at joint decisions. Groups are widely used across society, in contexts ranging from routine everyday decisions, such as choosing a shared activity with a friend, to high-stakes environments, such as medical teams determining treatment plans for critically ill patients. The widespread use of group decision-making is underpinned by the commonly held belief that "two heads are better than one" (2HBT1). However, empirical research indicates that this collaborative advantage is not always guaranteed (Blanchard et al., 2020). Instead, it is influenced by specific task conditions (Koriat, 2012, 2015), the use of confidence judgments as a metacognitive signal for correctness (Bahrami et al., 2010), and the quality and type of communication between group members (Bahrami et al., 2012; González-Romá & Hernández, 2014; Mahmoodi et al., 2013).

Metacognitive confidence refers to an individual's beliefs about the accuracy of their own decisions. It plays a crucial role in group decision-making because confidence serves as a socially interpretable cue: when members share and compare their confidence levels, this information can be used to guide their joint decision. Bahrami et al. (2010) showed that dyads achieve a 2HBT1 effect by adopting a confidence-weighted strategy that selects the response associated with higher confidence. This strategy is effective because higher confidence tends to be associated with a greater likelihood of correctness (Koriat, 2008; Yaniv, 1997). However, when the correlation between confidence and accuracy is weak, dyads struggle to identify which members' response is more likely to be correct and may fail to achieve a collective benefit (Blanchard et al., 2020). When this correlation is negative, the more confident member is more likely to be wrong, and dyads can perform worse than individuals

(Koriat, 2015). Metacognitive confidence is typically quantified using self-reported confidence ratings, which can be compared to objective performance to assess its calibration (i.e., overconfidence or underconfidence). This thesis extends prior work on dyadic decision-making by focusing on three key areas: 1) the role of task conditions; 2) the role of metacognitive confidence; and 3) the influence of communication.

The influence of asymmetric knowledge under dynamic task conditions (Study 1). Many real-world decision-making contexts involve dynamic environments, where task conditions change rapidly and members have access to different task-related information (information asymmetry). These scenarios require ongoing adaptation to maintain performance. Prior research suggests that effective communication and coordination are critical for groups to leverage their distributed knowledge and gain an advantage. However, most research on the 2HBT1 effect has focused on static tasks, and relatively little is known about how collaboration functions under unpredictable, changing conditions. Using a novel approach with a synthesis of qualitative and quantitative methodologies, Study 1 addresses this gap.

Collective intelligence and collective metacognition under static task conditions (Study 2). In contrast, static tasks involve stable operational environments and may be well-structured or ill-structured. Well-structured tasks have clearly defined goals and a single objective solution, whereas ill-structured tasks are open-ended with multiple solutions and pathways to completion (Laughlin, 2011). The construct of Collective Intelligence has been proposed to describe a group's general ability to perform across a broad range of tasks. Woolley et al. (2010) argued that collective intelligence is largely independent of individual cognitive ability and instead depends on factors such as social sensitivity, conversational turn-taking, and gender composition. However, recent research, including two meta-analyses, challenges this view, suggesting that individual cognitive ability and task characteristics (i.e., whether a task is well- or ill-structured) play a more important role in determining collective

outcomes. Metacognitive confidence, which is an individual's ability to monitor their own performance, may be particularly influential in well-structured tasks and has not been studied in this context. Using both novel methodological and analytical approaches, Study 2: 1) tested the robustness of the collective intelligence factor in dyads using well-structured tasks guided by the Cattell-Horn-Carroll (CHC) model of intelligence; 2) explored the relationship between dyadic collective intelligence and metacognitive confidence; and 3) identified the psychological characteristics of distinct dyad types.

The influence of trait confidence and communication type under static task conditions (Study 3). The confidence theory proposed by Bahrami et al. (2010) emphasizes the importance of confidence judgments in collaborative decision-making. According to this theory, group members rely on each other's expressed confidence as a cue for response accuracy, with higher confidence typically interpreted as a signal for correctness. Bang et al. (2017) extended this theory by showing that group members tend to align their confidence levels over time, a process referred to as confidence matching. Confidence is known to have the domain-general characteristics of a trait, and emerging evidence suggests that trait confidence may shape collective outcomes and possibly the development of confidence matching. However, the effects of trait confidence remain understudied. Using both novel methodological and analytical approaches, Study 3: 1) examined how trait confidence and type of communication moderate dyadic decision accuracy, decision confidence, and the alignment of decision confidence judgments; and 2) identified the psychological characteristics of distinct dyad types.

To better understand the conditions under which 2HBT1 emerges, this thesis investigated these three areas and explored how cognition, metacognition, and communication interact to shape dyadic outcomes for dynamic and static tasks. The use of different methodological and analytical approaches is critical for establishing the robustness and ecological validity of findings (e.g., Botvinik-Nezer et al., 2020). Convergence of results

from differing approaches provide compelling evidence that findings are not artifacts of the method and researcher bias but instead reflect naturally occurring processes and outcomes for dyads. The series of studies reported in this thesis were programmatically designed to employ distinct methodologies to enhance the evidence and overall conclusions.

1.2 Research Aims

The primary goal of this thesis was to investigate the psychological traits and communication factors that shape effective collaboration in dyads. Specifically, it examined how individual differences in cognitive ability, confidence, and communication influence dyadic outcomes across both dynamic and static task conditions.

This overarching aim was addressed through three focused research questions:

1. To investigate the 2HBT1 effect in a dynamic task involving an asymmetric distribution of information and changing operational conditions, and to examine how this effect relates to the quality and quantity of communication between dyad members.
2. To test whether collective intelligence, as defined by Woolley et al. (2010), emerges in dyads performing well-structured static tasks, and to evaluate the influences of individual cognitive ability, confidence, and communication structure on dyadic performance.
3. To assess how trait confidence and different methods of communication shape dyadic decision accuracy, decision confidence (task-specific confidence ratings), and the process of confidence matching in well-structured static tasks.

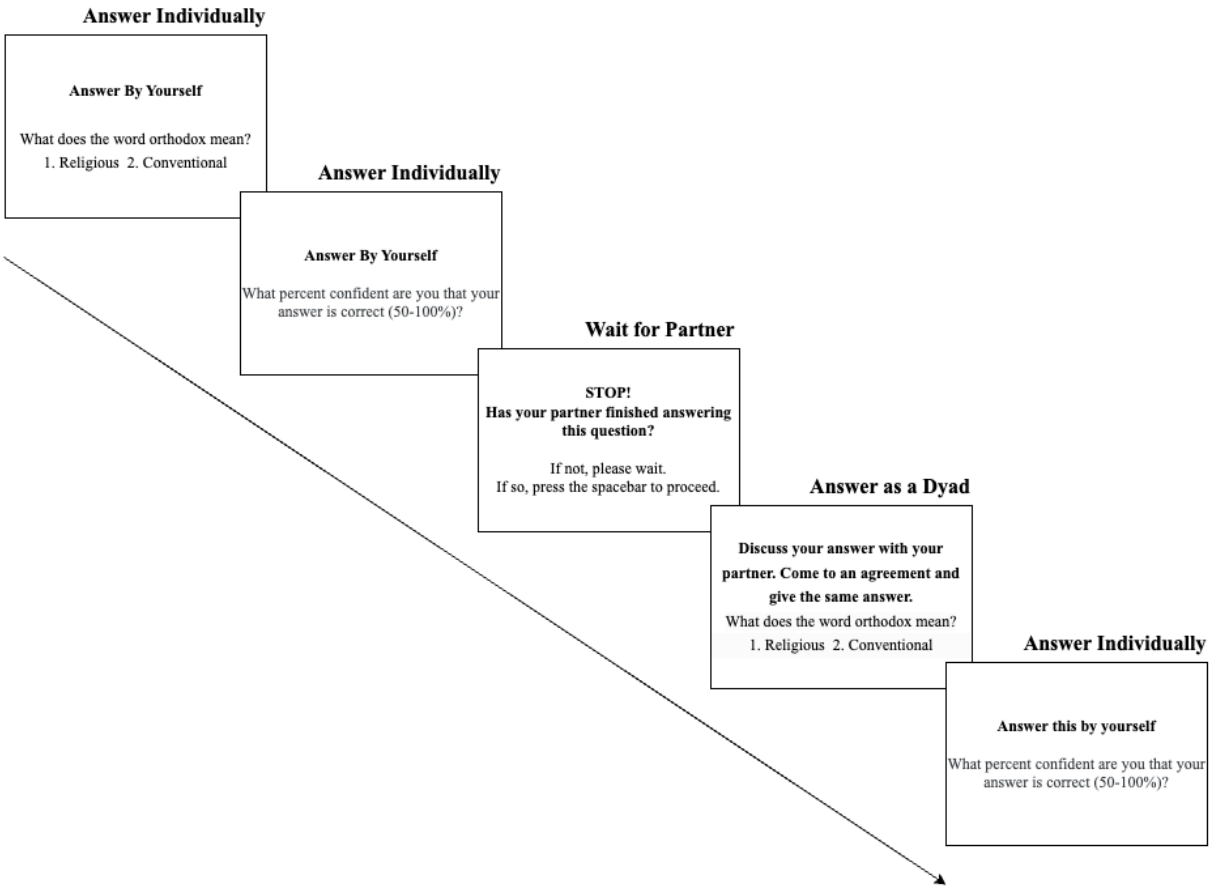
1.3 General Method

Study 1 (Chapter 2) employed a simulated driving task. Participants completed the task either individually (as drivers) or in dyads, where one member acted as a driver and the other as a navigator. The simulation included normal and foggy operational conditions, with fog appearing and disappearing suddenly to test adaptive performance. Dyads operated under asymmetric information due to their differing roles, requiring communication to form an accurate shared understanding of the task. Communication was measured in terms of quality and quantity to assess when the 2HBT1 effect emerged, and when it did not.

Study 2 (Chapter 3) examined dyadic performance on a set of well-structured tasks selected based on the CHC model of intelligence. Participants completed each item individually and then together as a dyad (this process is displayed in Figure 1.1). The study assessed the emergence of collective intelligence, as conceptualised by Woolley et al. (2010), and explored its relationship with individual intelligence, confidence, and communication. Latent Profile Analysis (LPA) was used to identify distinct dyadic profiles, highlighting heterogeneity in dyadic collaboration and performance.

Study 3 (Chapter 4) focused on the influence of trait confidence and communication mode. Dyads were composed of individuals identified through an extensive pre-screening process as high or low in trait confidence and were paired to form high, low, or mixed (one high and one low member) trait confidence dyads. They completed matched general knowledge tests as individuals and "dyads" under three different communication conditions: isolate (no interaction), passive (viewing partner's response and confidence rating), and active (verbal discussion). The study examined how trait confidence and type of communication influenced accuracy, decision confidence, and confidence matching. LPA was again used to classify distinct psychological profiles of dyadic collaboration.

Figure 1.1. *Order of Presentation for Study 2 and Study 3 Items (Active Communication Only)*



1.4 Thesis Structure

Chapter 1 provides an overview of the thesis, introducing its background, research aims, and summary of the empirical studies.

Chapter 2 presents Study 1: *Are two naïve and distributed heads better than one? Factors influencing the performance of teams in a challenging real-time task*. This study investigated the performance of dyads with asymmetrical roles on a dynamic task with changing operational conditions. The study developed a novel approach for measuring the quality of communication, offering insight into its relationship with dyadic performance under different operational conditions. This study has been published in a peer-reviewed journal (see the Related Scientific Output section).

Chapter 3 presents Study 2: *A Recipe for Dyadic Collective Intelligence for Well-Structured Tasks: Mix Equal Parts Cognitive Ability and Confidence Plus a Pinch of Social Sensitivity*. This study evaluates the validity of Woolley et al.'s (2010) collective intelligence construct using well-structured tasks guided by the CHC model. It also examines the relationship between collective intelligence and metacognitive confidence. The findings challenge the generalisability of earlier collective intelligence research and emphasize the importance of cognitive ability, confidence, and social sensitivity. This study has been published in a peer reviewed journal (see the Related Scientific Output section).

Chapter 4 presents Study 3: *How Trait Confidence and Communication Shape Dyadic Collaboration and Confidence Matching*. This study extends the findings of Study 2 by demonstrating how individual differences in trait confidence and different communication modes jointly shape dyadic decision accuracy, decision confidence, and confidence matching. It also extends prior work on confidence matching by showing that it occurs naturally during verbal interactions and that its relationship with accuracy may be context dependent. This study has been submitted for peer review with the journal *Cognitive Research: Principles and Implications* (see the Related Scientific Output section).

Chapter 5 integrates the findings from all three studies, discussing their contributions to the group decision-making literature. It highlights strengths, addresses limitations, describes theoretical and practical implications, and suggests directions for future research.

The appendices contain detailed supplementary materials for each of the studies.

Chapter 2: Study 1

Are Two Naïve and Distributed Heads Better than One? Factors Influencing the Performance of Asymmetrical Teams in a Challenging Real-time Task

The original manuscript for the study described in this chapter has been published in the journal *Frontiers in Psychology:* Blanchard, M. D., Kleitman, S., & Aidman, E. (2023). Are two naïve and distributed heads better than one? Factors influencing the performance of teams in a challenging real-time task. *Frontiers in Psychology*, *14*, 1042710. https://doi.org/10.3389/fpsyg.2023.1042710.

Minor adjustments have been made for this thesis to harmonise the language between the studies.

2.1 Introduction

Technological advances and situational necessities, such as the COVID-19 pandemic, have increased the capacity of geographically distributed individuals to collaborate (Bell & Kozlowski, 2002), particularly in dynamic environments. These environments are characterized by high demands on time and cognitive resources, ambiguous and rapidly changing information, and multiple interconnected decisions. Real-world scenarios often involve distributed groups performing high-stakes, mission-critical tasks with a deliberate division of labour across group members. In these roles, members typically have access to different information about the operating environment, and it is the pooling of these diverse perspectives that confers a performance advantage. Examples include air traffic controllers and pilots, forward observers directing air strikes, and operations centers coordinating emergency services.

Despite the growing importance of teamwork in dynamic environments, many studies examining the "two heads are better than one" effect (2HBT1) have focused on static tasks (e.g., Bahrami et al., 2010; Hill, 1982; Sniezek & Henry, 1989). These studies show that the 2HBT1 effect depends on task characteristics (e.g., Koriat, 2012a, 2015). However, there is a scarcity of research that has examined this effect using dynamic tasks, and to our knowledge,

none have done so using distributed groups with asymmetrical roles or under varying operational conditions (i.e., task characteristics). Given the increasing reliance on distributed groups, it is critical to examine how team asymmetry and task dynamics influence the 2HBT1 effect.

Communication is an essential process that groups use to share and process information (Hinsz, Tindale, & Volrath, 1997). Effective communication is associated with fewer errors across a broad range of tasks (e.g., Christensen et al., 2000; Donchin et al., 1995; Foushee, 1984; Helmreich et al., 1999; Kanki et al., 2010). Yet, as Marlow et al. (2018) note in their review, few studies have examined how communication affects asymmetrical team performance under different operational conditions. Moreover, communication quality is often assessed using post-task self-report questionnaires, which are prone to response distortion (e.g., Arnold & Feldman, 1981; Sackett, 1979).

To address these limitations, we developed a novel method for assessing communication quality during dynamic task performance. This method captures both the accuracy and timing of information shared between team members in real time. Unlike post-task self-reports, our approach provides a more behaviourally grounded measure of communication, allowing us to explore its relationship with team performance, both accuracy and speed, across varying operational conditions. If the 2HBT1 effect depends on the operational conditions of a dynamic task, then communication's impact on performance may also depend on these conditions. Our study aims to examine this relationship.

In the present study, we compared the performance of individuals and asymmetrical, distributed groups across two operational conditions in a dynamic task. We also developed a novel method for measuring communication quality to examine its relationship with team performance. We used naïve (also known as ad hoc) dyads, whose members had no prior experience working together. These types of groups are common in business and industry

(Devine et al., 1999; Sundstrom et al., 2000), but they also have a higher potential for failure compared to experienced groups. Understanding their performance outcomes, and the role of communication, in therefore critical.

2.1.1 Are 2HBT1 Under Different Operational Conditions?

Numerous studies have shown that two heads are more accurate than one on tasks with *static* environments (e.g., Bahrami et al., 2010; Hill, 1982; Laughlin, 2011; Sniezek & Henry, 1989; Tindale, 1989). However, this effect depends on the characteristics of the task. For example, Koriat (2012a, 2015) found that two heads are better than one for non-misleading cognitive tests items, but worse for misleading ones. Most of these studies have examined the 2HBT1 effect in symmetrical dyads, where both members perform the same role and are exposed to the same information (Hill, 1982; Laughlin, 2011; Sniezek & Henry, 1989; Tindale, 1989). For example, Koriat (2015) had dyads complete general knowledge questions together, with both members receiving identical information.

Several studies have introduced asymmetry by exposing them to different information while performing the same roles (Bahrami et al., 2010; Bahrami et al, 2012a, 2012b; Mahmoodi et al, 2015; Pescetelli et al, 2016). In Bahrami et al. (2010), participants performed a visual discrimination task where participants were briefly shown a pattern comprised of smaller circles and one target circle had a higher contrast than the others. Participants had to identify the target. During some trials, one member's view was degraded by added noise, creating an informational advantage for the other member. They found that the 2HBT1 effect persisted whether the information was shared or asymmetrically distributed. This is consistent with a key assumption of the 2HBT1 effect: dyads benefit from access to unique information that is distributed asymmetrically between members (Stasser & Titus, 1985, 1987). Our study builds on this logic by applying it to dynamic tasks with asymmetric team roles.

Several studies have extended the 2HBT1 effect to dynamic tasks involving symmetrical dyads (e.g., Abbott et al., 2021; Glynn & Henning, 2000; Räder et al., 2014; Shanks et al., 2013; Tolsgaard et al., 2015). For example, Glynn and Henning (2000) found that dyads were more accurate and faster than individuals on a dynamic teleoperation task that required participants to guide an object through a complex pathway. Other studies have explored how operational conditions, such as unexpected events or "roadblocks", can disrupt team functioning in dynamic environments. For example, Cooke and colleagues introduced changes (e.g., equipment failure, new target, enemy threats) into Uninhabited Aerial Vehicle (UAV) missions to examine teamwork processes such as situational awareness and coordination (Cooke et al., 2009; Gorman et al., 2005; Gorman et al., 2006; Gorman et al., 2010). These studies focused on *process* rather than performance outcomes, but their findings suggest that environmental disruptions may impair team functioning more than individual performance. Groups must engage in additional communication to reestablish shared understanding and recover performance. The increased communication demands that arise when adapting to changing task conditions may require greater cognitive and temporal resources (MacMillan et al., 2004), which can impair performance, especially for naïve dyads. When the task environment changes, team members must divide their cognitive resources between executing their role and managing communication. This increased cognitive load may harm team performance, and under certain conditions, dyads may underperform compared to individuals, potentially reversing the 2HBT1 effect.

In the present study, participants completed a dynamic driving simulation under two operational conditions (normal and fog) as either individuals or as members of asymmetrical, distributed dyads. Dyads were composed of a driver and a navigator. The driver's task was to navigate a vehicle as quickly and safely as possible toward a target destination indicated by arrows. The navigator's task was to monitor the environment and provide information or instruction to support the driver's goals. The fog condition introduced an unexpected

environmental change that increased the cognitive workload of the driver, disrupting performance and required adaptation to recover. In the normal condition, both team members experienced the same conditions and had access to similar information. In the fog condition, only the driver's visibility was impaired, conferring an informational advantage to the navigator. Performance was assessed using two metrics: the driver's accuracy (collisions) and speed (as a proxy for time). Given that humans have limited cognitive capacity, we did not expect the 2HBT1 effect to emerge across both conditions and both performance metrics. Rather, we expected that the effects of teamwork would depend on communication

2.1.2 Team Communication

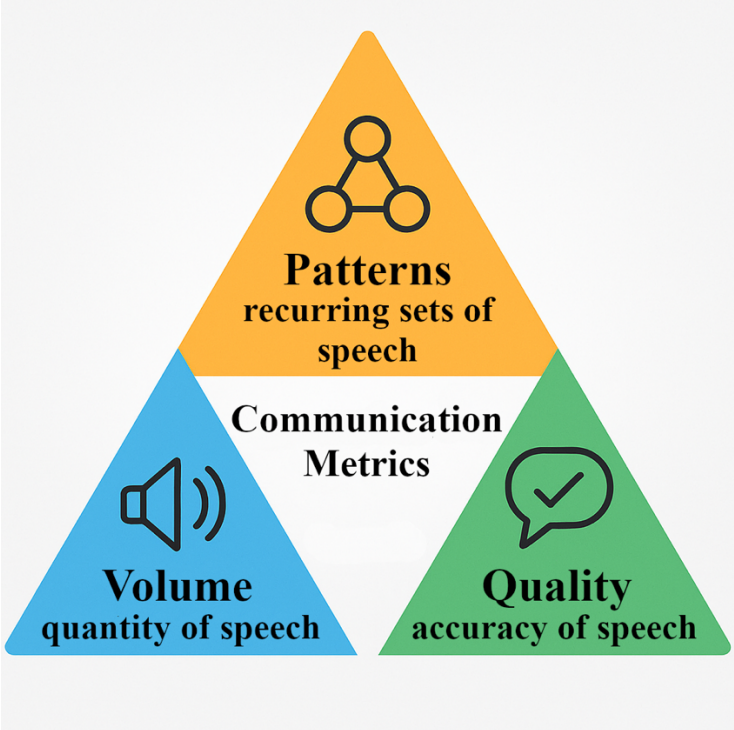
Communication is an essential process when working in a group (Keyton et al., 2010) and has consistently been shown to distinguish high- and low-performing groups (Cooke et al., 2007). Broadly, team communication involves the exchange of information among team members (Adams, 2007), and groups have been conceptualised as information-processing units (Hinsz et al., 1997; Tindale & Kameda, 2000). Communication plays a central role in many models of team performance, underpinning both team processes (e.g., coordination) and emergent states (e.g., shared mental models and team cognition). It also has both direct and indirect relationships with team performance (e.g., Marks et al., 2001; Mathieu et al., 2008; Ilgen et al., 2005; Kozlowski & Bell, 2013; Kozlowski & Klein, 2000).

Through communication, groups share task-relevant information (Salas et al., 2005) and situational factors (MacMillan et al., 2004) to develop a shared understanding (Endsley, 1988; Rouse & Morris, 1986), resolve misunderstandings (Fletcher & Major, 2006), coordinate actions, and formulate strategies (Marks et al., 2001). Despite its critical role, how best to measure team communication remains a topic of debate.

2.1.2.1 Measurement

Common metrics used to assess communication include volume, quality, content, and patterns. Each captures distinct properties of communication and relates to team performance in different ways (see Figure 2.1).

Figure 2.1. *Common Communication Metrics*



Volume refers to the duration of speech and frequency of speaking turns (Bunderson & Sutcliffe, 2003; Woolley et al., 2010). While easy to measure, volume-based metrics ignore the content and accuracy of what is communicated. A recent meta-analysis (Marlow et al., 2018) found that measures of communication frequency had weaker associations with team performance than communication quality, which are more difficult to capture objectively. Volume may also serve as a proxy for cognitive load, as communication is cognitively demanding (MacMillan et al., 2004). In this study, we treated volume as a baseline rather than a primary metric.

Quality refers to the clarity, accuracy, and timeliness of information shared between members. Higher quality communication is typically associated with better performance

(González-Romá & Hernández, 2014; Hirst & Mann, 2004). However, most assessments of communication quality rely on self-report questionnaires completed after the task, where participants rate overall communication quality on a Likert scale (Marlow et al., 2018). While efficient, this method has several limitations: it does not consider the content of communication or its objective accuracy, and it is susceptible to biases, such as recall and social desirability (e.g., Arnold & Feldman, 1981; Sackett, 1979). Additionally, this approach treats communication as static, despite evidence that it is dynamic and dependent on the operational environment (Cooke et al., 2009). Smith-Jentsch (2009) noted a lack of research assessing the accuracy of knowledge quality shared between teammates using more objective measures. Our study addresses this gap by developing a novel, behaviour-based method.

We developed and applied a *manual content analysis* approach to assess communication quality as an alternative or supplement to self-reports. It requires a researcher to select or develop a classification scheme to reduce the complexity of communication data to several categories that represent both the linguistic content of a team's interactions and cognitive processes (e.g., knowledge sharing, information processing, and planning). For example, Bowers et al. (1998) coded speech into seven categories, such as uncertainty, action, acknowledgement, and non-task related statements. They found that high-performing groups acknowledged or responded more frequently to task-relevant verbal acts and engaged less in non-task related talk. A limitation of these approaches is the large amount of time and resources involved in developing and validating coding schemes, and the potential lack of cross-validation for behaviour classification.

In the present study, we used manual content analysis to quantify communication quality. We defined high-quality communication as speech that was both accurate in the current situation and delivered at a time when the teammate could act on it. This approach also allowed us to examine communication patterns, which are discussed in the next section. All recordings were cross validated prior to inclusion in the main analyses.

Communication patterns refer to recurring sets of verbal behaviours that emerge as people work together (e.g., Bowers et al., 1998; Fischer et al., 2007; Gorman et al., 2012). For example, Stachowski et al. (2009) found that high-performing groups engaged in fewer, shorter, and less complex patterns of communication than lower-performing groups. However, many existing approaches to identifying patterns tend to ignore the semantic content, thereby missing key aspects of communication quality (Wildman et al., 2014).

In our study, we applied this approach in a novel way to identify stable and recurring speech patterns that represented the quality of communication within each operational condition of the dynamic driving simulation. By including both structure and content, we aimed to better capture how communication evolves in response to changing task demands.

2.1.2.2 The Cost of Communication

The 2HBT1 effect suggests that dyads typically outperform individuals across a range of tasks. However, not all dyads perform better. One reason may be that dyads have limited cognitive resources, and communication is cognitively and temporally demanding (MacMillan et al., 2004). Efficient communication is therefore vital to team performance. MacMillan and colleagues observed that the more speech required to convey a message, the greater the cognitive overhead in terms of both time and mental effort. Depending on the cognitive resources available, some dyads may thrive while others decline under increased communication demands. This aligns with research on information overload, which suggests that higher communication frequency often contains more irrelevant or distracting information (See Edmunds & Morris, 2000 for review). It is also consistent with research on mobile phone use while driving, which shows that communication while operating a vehicle slows reaction times, increases caution, and decreases speed (e.g., Alm & Nilsson, 1994; Cooper et al., 2003; Haigney et al., 2000; Lamble et al., 1999; McKnight et al., 1993).

Extending these findings to our dynamic driving simulation, we expected that increased communication would negatively impact driving speed, and potentially accuracy. Specifically, when the task shifted from normal to fog conditions, drivers may reduce their speed to compensate for the additional cognitive load and this reduction may maintain accuracy (avoiding collisions). Key implications for our study and our novel communication metrics are discussed in the next section.

2.1.2.3 Team Communication in the Present Study

To improve dyadic performance and the efficient allocation of organisational resources, it is important to understand when communication is helpful and when it may harm outcomes, particularly in dynamic environments that involve asymmetrical, distributed dyads. In this study, we expected that dyads would demonstrate higher accuracy than individuals during normal conditions but not during the fog condition, when task demands and the need for communication increase. We also expected that dyads would drive more slowly than individuals under both conditions. These predictions were based on three main assumptions: 1) communication is cognitively demanding and people tend to reduce speed when communicating; 2) during the normal condition, the driver's cognitive load would be moderate, leaving enough capacity to attend to communication without sacrificing accuracy; and 3) during the fog condition, the driver's cognitive load would be higher and attending to communication would require diverting cognitive resources away from driving which would reduce accuracy. Within dyads, we expected that only the driver's performance would be affected by communication during the fog condition, as the navigator's role remained relatively constant across both operational conditions. That is, the navigator's cognitive load should remain lower than the driver's during normal conditions and should not increase substantially during the fog, since their view of the operational environment remained unchanged.

Many previous studies fail to assess the psychometric properties of communication variables (e.g., internal consistency; Wildman et al., 2014). This is a critical omission, as these analyses are necessary to evaluate the reliability and validity of communication measures and to investigate whether stable latent structures of communication exist. To address this, we developed a novel method for assessing communication quality during a dynamic driving simulation that used multiple iterations to allow psychometric evaluation. We then used our communication quality metrics, alongside traditional volume-based metrics, to predict team performance on accuracy (collisions) and speed. We also compared their relative predictive power while controlling for other theoretically relevant individual differences variables. We expected that our novel measures of communication would be a stronger predictor of accuracy, whereas volume of communication metrics would be a stronger predictor of speed.

2.1.3 Simulation-Based Assessment

In the present study, we employed a high-fidelity driving simulation that we developed in previous research (Kleitman et al., 2022; Kleitman et al., 2020). Simulations are open-ended, rule-based environments that allow participants to engage with artificial problems and generate quantifiable outcomes (Salen & Zimmerman, 2004). They are well suited for assessing complex skills and behaviours in dynamic environments because, compared with traditional assessment tools, simulations offer “free play” experiences (Mislevy, 2013; Shute & Ke, 2012); and better reflect the physical and psychological characteristics of real-world tasks (Beaubien & Baker, 2004; Bowers & Jentsch, 2001). Driving is a complex task that requires vision, visual perception, physical control, emotional control, information processing, and executive functions (Anstey et al., 2005; Asimakopulos et al., 2012, Mathias & Lucas, 2009). Because individuals differ in these abilities, we also measured a range of psychological variables associated with driving performance and collective performance.

2.1.4 The Role of Individual Differences

We controlled for the following individual differences variables when predicting team performance during the driving simulation. We also accounted for prior driving and gaming experience.

Executive functions are a set of mental processes involved in maintaining focus, resisting distractions, and adapting to changing task demands (Diamond, 2013). Three core executive functions are most relevant to dynamic tasks like our simulation: 1) inhibitory control is the ability to suppress impulsive or premature responses (De Jong et al., 1995); 2) working memory is the capacity to hold and manipulate information in mind which enables pattern recognition (Baddeley, 1992); and 3) cognitive flexibility is the ability to shift strategies and adapt to new rules or task demands (Davidson et al., 2006). All three are essential in dynamic tasks where environmental conditions can change rapidly and unpredictably.

Cognitive ability refers to one's capacity to process information and to implement actions to achieve goals (Carroll, 1993). Two key types of cognitive abilities are fluid intelligence and crystallised intelligence (Cattell, 1971, 1987). Fluid intelligence refers to the ability to reason, problem solve, and process novel information in real time. Crystallised intelligence refers to knowledge acquired through experience and learning. Given the complexity and novelty of our driving simulation, we focused on fluid intelligence, which is more relevant to the demands of our dynamic task.

Personality has been linked to team performance, particularly the Big Five traits. Meta-analyses show that Agreeableness, Conscientiousness, and Openness to Experience are positively associated with team performance (Bell, 2007; Peeters et al., 2006). In the context of driving, low Conscientiousness has been linked to a greater likelihood of traffic accidents (Arthur & Graziano, 1996; Sümer et al., 2005), while neuroticism is associated with poorer

task performance when under stress (Schneider, 2004). Dynamic simulations, such as ours, can be stressful experiences because the environment and demands of the task may rapidly change, thus, Neuroticism, which indicates one's emotional stability, may also be relevant to our task.

Demographic characteristics also matter. For example, the proportion of females has also been shown to relate with team performance across a range of static tasks. Woolley et al. (2010) observed that groups with a higher proportion of females tended to perform better than those with a lower proportion of females.

2.1.5 The Present Study

Participants completed a dynamic driving simulation under two conditions, each with two levels: grouping (individual vs. dyad) and operational condition (normal vs. fog). Grouping was a between-subjects factor where participants completed the simulation either individually or as asymmetrical, distributed dyads consisting of a driver and a navigator. The operational condition was a within-subjects factor where all participants were exposed to both the normal and fog conditions during each lap of the driving simulation. In the normal condition, both the driver and navigator had high visibility and access to similar information about the environment. In the fog condition, a sudden onset of dense fog that reduced the driver's visibility but did not affect navigator's view which included an aerial perspective via UAV. As a result, the navigator had an informational advantage in the fog condition. All other environmental characteristics remained consistent across both conditions.

We recorded and coded team communication during the simulation and linked it to two performance metrics: accuracy (collisions) and speed (a proxy for time). The contrast between operational conditions allowed us to examine whether the 2HBT1 effect, and the relationship between communication and performance, depended on task characteristics, particularly when one member held an informational advantage. We also developed a novel

coding system to quantify the quality of knowledge shared between dyad members, focusing on accuracy and the appropriateness of timing. Each speaking turn was coded into one of five behavioural categories that captured team cognition. The categories were: 1) observation; 2) command instruction; 3) inquiry; 4) redundant; and 5) frustration. For the navigator's speech, we further assessed whether observations and command instructions were helpful (accurate and well-timed) or harmful (inaccurate and/or ill-timed). For each team, we calculated the frequency of each communication category during each operational condition and each lap. Details of the coding system are provided in the Method section.

To identify stable patterns of communication, we conducted exploratory factor analysis (EFA) on the coded communication categories. While we did not formulate specific hypotheses about the factor structure due to the novelty of the approach, we formulated exploratory predictions. Our communication coding targeted helpful (accurate and well-timed) and harmful (inaccurate and ill-timed) communication. Thus, we expected to extract two factors that captured helpful and harmful communication patterns. We then examined whether these patterns predicted dyadic performance using hierarchical regression models, separately for each operational condition.

The overarching aim of the present study was to better understand performance in asymmetrical, distributed dyads operating under different operational conditions. We used a mixed design (grouping: between-subjects; operational condition: within-subjects) to address following aims and hypotheses.

Aim1: To investigate whether the 2HBT1 effect emerges across different operational conditions in a dynamic task.

We predicted that dyads would be more accurate (fewer collisions) than individuals during the normal condition, but not during the fog condition, when communication demands

are higher. We also expected that dyads would drive slower than individuals in both normal and fog conditions due to the cognitive cost of communication.

Hypothesis 1a: Dyads would have fewer collisions than individuals during the normal condition.

Hypothesis 1b: Dyads would drive slower than individuals in the normal condition.

Hypothesis 1c: There would be no difference in collisions between dyads and individuals during the fog condition.

Hypothesis 1d: Dyads would drive slower than individuals during the fog condition.

Aim 2: To examine the relationship between communication quality and dyadic performance across the two operational conditions, after controlling for theoretically important individual differences variables.

We expected that helpful communication patterns would be associated with higher performance (fewer collisions and slower speed to maintain control), while harmful patterns would be associated with poorer performance in both operational conditions (higher collisions and slower speed to maintain control). Given the cognitive overhead associated with communication (MacMillan et al., 2004) and that drivers tend to go slower when communicating (Haigney et al., 2000), we expected that both helpful and harmful communication would be associated with lower speed in both operational conditions. That is, regardless of the quality of communication, higher levels of communication would be detrimental to speed.

Hypothesis 2a: Helpful communication would negatively predict collisions in the normal and fog conditions (i.e., dyads would have fewer collisions when communicating more effectively).

Hypothesis 2b: Helpful communication would negatively predict speed in the normal and fog conditions (i.e., dyads would drive slower when communicating more effectively).

Hypothesis 2c: Harmful communication would positively predict collisions in the normal and fog conditions.

Hypothesis 2d: Harmful communication would negatively predict speed in the normal and fog conditions.

Aim 3: To evaluate whether our novel communication quality metrics were stronger predictors of dyadic performance than traditional volume-based measures (i.e., duration of speech and number of speaking turns). We expected that the content of communication would more strongly relate to accuracy, while the volume of communication would better predict speed, reflecting the cognitive cost of communication.

Hypothesis 3a: Communication quality metrics would be stronger predictors of collisions than communication volume.

Hypothesis 3b: Communication volume metrics would be stronger predictors of speed than communication quality.

2.1.6 Statistical Analyses

To identify stable patterns of communication quality, we conducted an EFA on the coded communication behaviours using our novel classification system. To test hypotheses 1a-1d, we conducted a series of two-way mixed design ANOVAs examining differences in performance between grouping conditions (between subjects: individuals vs. dyads) and operational conditions (within-subjects: normal vs. fog) on our two outcomes: accuracy (collisions) and speed. To test hypotheses 2a-2d, we used a series of hierarchical regression

analyses to examine the relationship between the extracted communication quality factors and dyadic performance (collisions and speed), after controlling for relevant individual differences variables. To test hypotheses 3a and 3b, we repeated the hierarchical regression analyses from Aim 2 but replaced the communication quality factors with traditional volume-based measures of communication. We then qualitatively compared the standardised regression coefficients to assess whether communication quality or communication volume served as a stronger predictor of each performance metric. All analyses were conducted using the R programming language.

2.2 Method

2.2.1 Participants

In return for partial course credit, 316 Australian undergraduate psychology students completed the study either alone or as a dyad (213 females, 103 males, mean age = 19.80, SD = 4.13). A total of 22 participants (12 individuals and 5 dyads) were excluded from analyses (see Appendix A for details). The final sample included 294 participants (196 females, 98 males, mean age = 19.80, SD = 4.12). 134 participants completed the driving simulation as individuals (87 females, 47 males, mean age = 19.80, SD = 3.35) and 160 participants completed the simulation as 80 dyads (109 females, 51 males, mean age = 19.70, SD = 4.69). Although the two groups were unequal in size, both were deemed adequate for the planned analyses. We followed the general rule of at least 10 observations per predictor (Tabachnick & Fidell, 2007), and our sample compared favourably with prior research on dyads, which used samples of between 15 and 43 dyads (Bahrami et al., 2010; Gorman et al., 2005; Glynn & Henning, 2000; Koriat, 2015; Sniezek & Henry, 1989).

2.2.2 Measures

Raven's Advanced Progressive Matrices (RAPM; Raven, 1938-65). This test is a measure of abstract reasoning. Each trial presented a 3x3 matrix of abstract figures following

a pattern horizontally and/or vertically. The bottom right figure was blank, and participants decided which of eight alternatives completed the pattern. A 20-item version (of 36) was used to reduce testing time. It was used to measure both abstract reasoning and cognitive confidence. Internal consistency estimates are acceptable to good for accuracy (Cronbach's Alpha = .68 - .86) and excellent for cognitive confidence (Cronbach's Alpha = .84 - .96; Blanchard et al., 2020; Jackson & Kleitman, 2014; Jackson et al., 2016a, 2016b; Jackson et al., 2017).

Random Number-Letter Switching Test (Monsell, 2003). This executive function task is a measure of cognitive flexibility. Participants were shown a cue (e.g., "letter" or "number") followed by a stimulus pair (e.g., "A6") and responded based on the cue. For example, if the instruction was "letter" and "A6" appeared on the screen then participants determined whether the letter on the screen was a vowel or consonant. Trials were classified as **repeat** or **switch** depending on whether the cue remained the same from one trial to another (repeat) or changed (switch). The task included 16 practice trials and 72 test trials. Outcome variables include repeat and switch error rates, response times when the response was correct, and switch cost (average switch response time minus average repeat response time).

Flanker test (Eriksen & Eriksen, 1974). This executive function task assessed inhibitory control. Each trial presented five horizontally aligned arrows with the centre arrow either pointing left or right. The other four arrows either pointed in the same direction (congruent trial) or the opposite direction (incongruent trial) as the centre arrow. Participants indicated whether the centre arrow pointed left or right. The task included 30 practice trials with feedback and 100 test trials without feedback. Outcome variables include error rates for congruent and incongruent trials, response times when the response was correct for both trial types, and inhibitory cost (average incongruent response time minus average congruent response time).

Running Letter Span (Broadway & Engle, 2010; Pollack et al., 1959). This task is a measure of working memory. Participants were instructed to recall the last *n* letters (range: 3-7) without knowing the total sequence length (range: 5-9). A sequence of individually appearing letters then flashed onto their computer screen. For example, they were instructed to remember the last 2 letters and the sequence "X Y T R S" appeared then the answer was "R S". The task contained 5 practice trials with feedback and 15 test trials without feedback. Internal consistency estimates are excellent for accuracy on this test (Cronbach's Alpha = .85; Broadway & Engle, 2010)

Mini International Personality Item Pool (IPIP; Donnellan et al., 2006). This 20-item questionnaire is a measure of the Big Five personality traits: Agreeableness, Conscientiousness, Extraversion, Intellect, and Neuroticism. Participants rated statements, such as "Am the life of the party", using a five-point Likert scale from (1) *Very inaccurate* to (5) *Very accurate*. There were four statements for each personality dimension. Each dimension has demonstrated acceptable internal consistency (Cronbach's Alpha range = .65 - .77; Donnellan et al., 2006).

NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988). This 6-item questionnaire is a measure of subjective workload during a dynamic task. It was administered immediately after the driving simulation to assess participants' confidence in their performance during the simulation. Only one item was relevant for this purpose, "How successful do you think you were in accomplishing the goals of the task?", Responses were recorded on a 7-point scale ranging from (1) very low to (7) very high. Test/retest reliability of the NASA-TLX has been shown to be good (.83; Hart & Staveland, 1988). As only a single item was used in our study, internal consistency could not be estimated.

2.2.3 Driving Simulation

We employed a driving simulation developed in prior work (Kleitman et al., 2022; Kleitman et al., 2020). Participants completed an emergency driving course in an urban environment, either as individuals or asymmetrical dyads (driver and navigator). Dyad members did not know each other prior to participation. The simulation goal was to deliver essential supplies as quickly and safely as possible while following directional arrows. Drivers could disregard traffic rules (e.g., traffic lights and line markings on the road), but they were encouraged to minimize collisions and maximize speed. Navigators operated a UAV and had access to a birds-eye view of the driving environment, in addition to seeing the driver's perspective (see Figure 2.2). Thus, navigators had access to shared and unique information relative to the driver. Their role was to observe the dynamic conditions experienced by drivers and communicate relevant instructions or warnings to them. For example, they could notify the driver when it was safe to use an oncoming traffic lane or warn them to reduce speed if they noticed traffic congestion ahead.

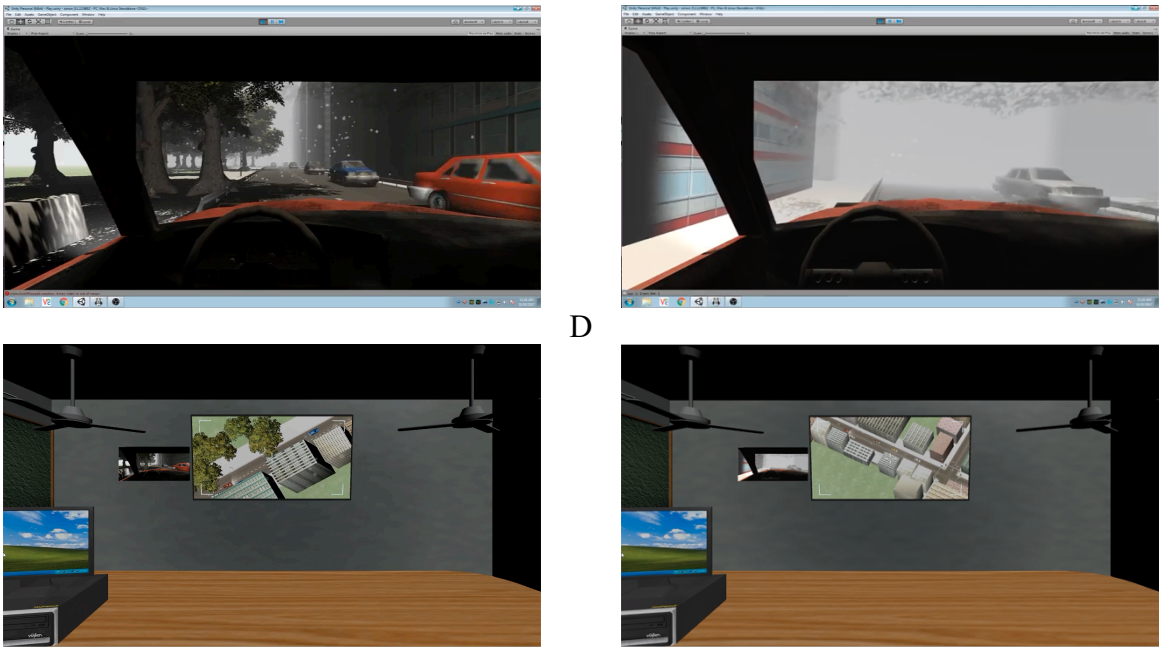
Participants in dyads were seated in different rooms (consistent with the scenario) so they could not see each other but they could freely communicate via headsets using a push-to-talk intercom system. This allowed us to record the speaker and receiver's identities, and the duration, timing, and content of speech. The simulation comprised five unique laps, that started and ended at the same location. Each lap included both a normal and fog condition. The fog condition introduced a sudden reduction in visibility for the driver but not the navigator. Both conditions were designed to last approximately the same amount of time. This structure allowed for psychometric assessment of both the communication and performance metrics.

Performance on the driving simulation was evaluated based on the driver's collisions and speed. Collisions were defined as any instance where the driver's vehicle came into contact with any other vehicle or object in the environment. Speed was recorded in notional

kilometres per hour and used as a proxy for time. The frequency of each communication type was recorded, and the method is described in the next section.

Figure 2.2. The Driver's and Navigator's Screens During the Normal Condition (A and C) and Fog Condition (B and D)

A B



C D

2.2.4 Quantifying Team Communication

2.2.4.1 Categories of Communication Behaviour

To assess team communication, we used a manual coding approach consistent with prior research (e.g., Bowers et al., 1998; Krippendorff, 2004; Predmore, 1991). Each speaking turn was assigned to one of several role contingent categories. To define the categories, we first identified unique communication behaviours that captured team cognition (Salas et al., 2007). Specifically, we targeted speech that indicated knowledge updating, information processing, and reduced cognitive capacity (see Table 2.1 for examples). Effectively performing one's role required different types of communication behaviour for drivers and navigators. Drivers primarily updated shared knowledge by providing situational feedback. Navigators were responsible for monitoring, instructing, and updating shared knowledge while coordinating the driver's actions. Because driver's speech had minimal

impact on their own performance, and because drivers relied heavily on the navigator's guidance, our assessment of communication quality focused on navigator speech.

Navigator speech was coded into four main categories: observation, command instruction, inquiry, and redundant (see Table 2.1 for definitions and examples). For observation and command instruction, we further assessed quality by coding them as either helpful (accurate and well-timed) or harmful (inaccurate and/or ill-timed). Helpful communication reflected accurate information delivered while it was actionable by the driver. Harmful communication contained misinformation or was poorly timed, diminishing its utility.

Driver speech was coded into three categories: command instruction/observation, inquiry, and frustration (see Table 2.1 for examples). We combined command instruction and observation into a single category after a preliminary analysis of the driving simulation recordings. These instances appeared to be of the same nature and the content of a driver's speech did not appear to impact their own performance, so we did not assess the quality of these speaking turns. Frustration was included to capture signs of cognitive overload such as stress, which has been linked to reduced communication and impaired performance (Cohen & Cohen, 1980; Driskell & Johnston, 1998; Driskell et al., 1999; Ellis, 2006; Gladstein & Reilly, 1985; Zheng et al., 2012).

Table 2.1. *The Communication Coding System*

Category of communication	Component of team cognition	Definition	Example
Navigator			
Helpful observation	Team knowledge building: a transference of *accurate* individual knowledge to team knowledge.	The information is appropriate and accurate for the driver's current situation.	"There's a lot of traffic up ahead." "There's a right turn coming up in 10 seconds."
Harmful observation	Team knowledge building: a transference of	The information is not appropriate and/or inaccurate for	"There's no traffic up ahead." [When the road ahead has numerous vehicles.]

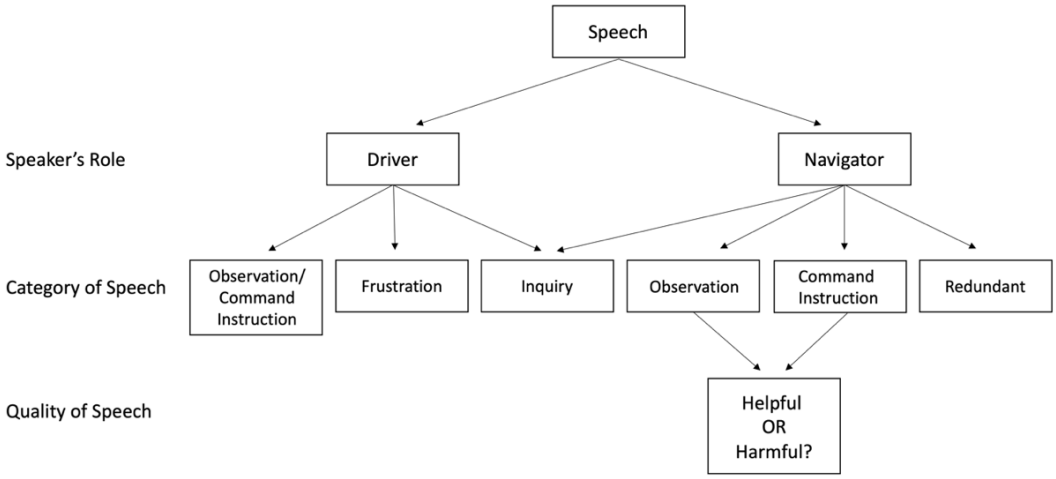
	inaccurate individual knowledge to team knowledge.	the driver's current situation.	There's a right turn coming up ahead." [When the arrows at the upcoming intersection are pointing left.]
Helpful command instruction	Performance monitoring: the navigator *accurately* advises the driver about a future course of action.	The instruction is appropriate and accurate for the driver's current situation.	"Switch lanes to overtake the car in front." "Wait for the traffic light to turn green because there are cars in front and behind your vehicle."
Harmful command instruction	Performance monitoring: the navigator inaccurately advises the driver about a future course of action.	The instruction is not appropriate and/or inaccurate for the driver's current situation.	"Switch lanes to overtake the car in front." [When the other lane contains incoming traffic.] "Reverse then switch to the other side of the road to avoid the traffic at the red light." [When the driver's car is stationary at a red light and surrounded by traffic ahead and behind.]
Inquiry	Team knowledge processing.	Seeking new information or clarification of existing knowledge about the task.	"What do you see ahead of you?" "Why have you stopped your car?"
Redundant	Reduced cognitive resources (e.g., Increased communication overhead).	The information or instruction is irrelevant for the driver's current situation or task.	"Move into the other lane." [after the driver has already changed lanes.] "I am hungry."
Driver			
Command instruction/ Observation	Team knowledge building.	The information or instruction relates to their current situation or the task.	"There's a red light ahead of me." [When asked why they stopped their car.] "I'm currently driving through fog so my visibility is low."
Inquiry	Team knowledge processing.	Seeking new information or clarification of existing knowledge about the task.	"Can you tell me when there are a lot of cars ahead?" Can you tell me what is around me now?" [When they have lost visibility during the fog event.]
Frustration	Cognitive overload.	Driver produces an audible expression of frustration.	"Oh no!" [As the driver's car collides into another vehicle.] An audible sigh when the driver is stationary and stuck in dense traffic.

2.2.4.2 Coding communication

Four independent raters, who were blind to the study's hypotheses, coded all simulation recordings. See Figure 2.3 for a visual representation of the process of coding each speaking turn. We developed a custom browser-based coding application (see Figure 2.4) using JavaScript, HTML, and CSS to facilitate consistency and precision. The tool allowed raters to watch each session and assign communication codes to individual speaking turns in real time using keyboard shortcuts. Raters could pause, rewind, and fast-forward as needed. Timestamps were recorded to align communication events with specific laps and operational conditions.

Prior to coding, all raters completed an extensive training session which consisted of defining each category of communication, providing examples of each category, demonstrating how to use the coding software, and coding a segment of a recording together as a group. To assess interrater reliability, all four raters independently coded the same 37-minute recording, which contained 539 speaking turns. Fleiss' (1971) Kappa revealed high agreement ($\kappa = .83$). When discrepancies occurred, the raters were instructed on how to code the communication according to the definitions. The remaining recordings were then divided among the raters and coded independently. After initial coding, an independent quality control review was conducted. A trained reviewer examined each recording to ensure each assigned category aligned with their definitions. Discrepancies were resolved by realigning codes as needed, according to the original operational definitions (See Table 2.1).

Figure 2.3. *Process of Coding Each Speaking Turn During the Driving Simulation*



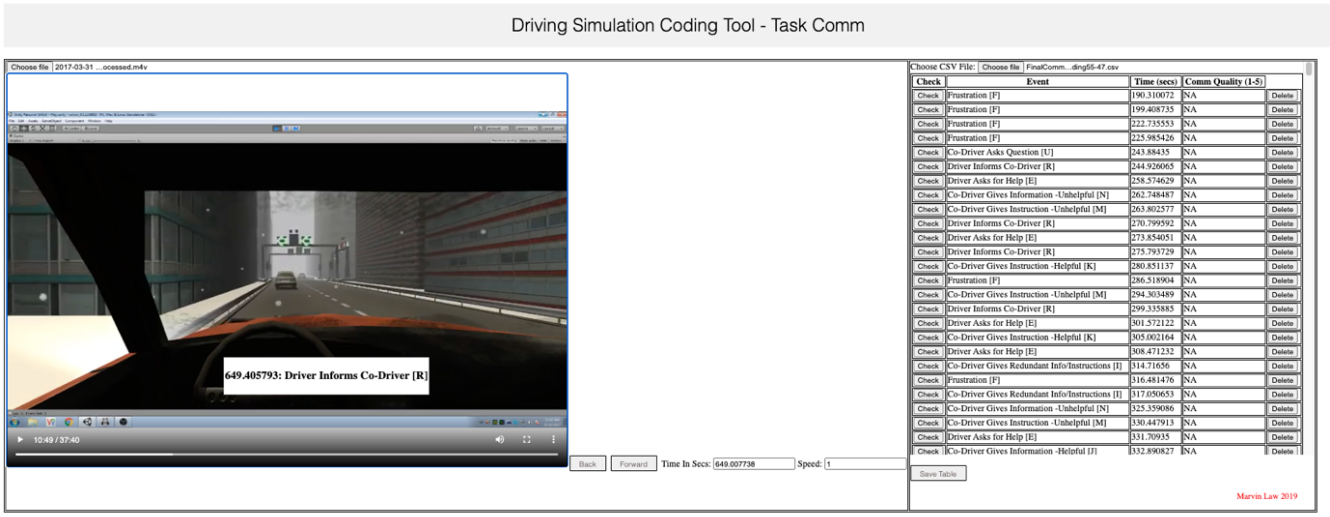
2.2.5 Procedure

Up to four participants completed the two-hour study at a time in university computer labs. Upon arrival, participants were randomly assigned to either the individual or dyad condition and followed the same protocol and task order regardless of condition. For participants assigned to dyads, the **driver** and **navigator** were seated in separate rooms to simulate distributed teamwork. They communicated via **headsets** using a **push-to-talk intercom** system, which allowed speech to be timestamped and recorded.

First, participants completed background questionnaires, which assessed demographic information (e.g., age and sex), experience with driving, gaming and simulations, and their susceptibility to motion sickness. Next, they completed the driving simulation, followed by the NASA-TLX, and then the cognitive and personality measures in the order described in the Measures section. The driving simulation was completed either individually or in dyads

depending on the assigned condition. Ethics approval was granted by the Australian Defence Science and Technology Group's Low Risk Ethics Panel (Protocol Number LD14-16)[1].

Figure 2.4. *The Communication Coding Application*



2.3 Results

Prior to testing our hypotheses, we examined descriptive statistics and internal consistency for the simulation derived measures of performance, individual difference variables, and the coded communication metrics. This allowed us to check that scores on each variable were as expected, assess reliability, and verify that individuals and dyads were similar on the control measures. Omega total (McDonald, 1999) was used to measure internal consistency because we assumed unidimensionality but not tau-equivalence for collisions and speed during the driving simulation.

2.3.1 Descriptive Statistics

2.3.1.1 Simulation Derived Performance Metrics

Frequency distributions for collisions overall and speed overall (across all laps and operational conditions) for individuals and dyads are displayed in Figure 2.5. Collisions

[1] The tasks were administered as part of a broader protocol examining a set of hypotheses regarding psychological resilience and adaptability in a high-fidelity simulation. A number of additional measures were captured to examine the hypotheses of interest and are outside of the scope of this paper.

overall were positively skewed, whereas speed overall was normally distributed for both individuals and dyads. These distribution patterns were consistent across both operational conditions (see Figure A1 and A2 in Appendix A). Table 2.2 reports the descriptive statistics, and internal consistency estimates for collisions and speed, separately for individuals and dyads under each condition.

Figure 2.5. *Frequency Distributions for Individual Collisions (A) and Speed (B) Overall and Dyad Collisions (C) and Speed (D) Overall*

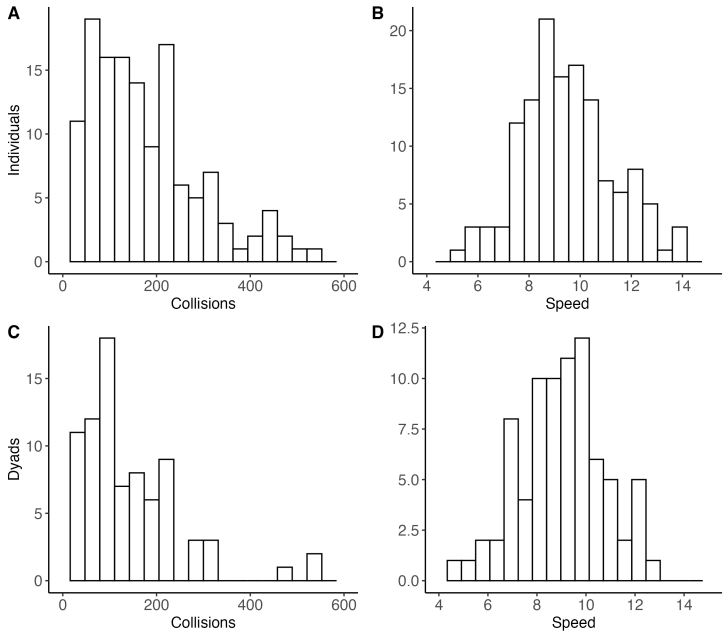


Table 2.2. *Descriptive Statistics and Internal Consistency Estimates for Simulation Performance (N=80)*

		Individuals			Dyads		
		ω_t	Mean	SD	ω_t	Mean	SD
Collisions	Normal	.81	100.29	73.12	.88	76.79	69.18
	Fog	.82	76.65	50.99	.73	67.20	46.16
Speed	Normal	.86	9.60	1.85	.88	9.20	1.78
	Fog	.82	9.45	1.87	.83	8.79	1.78

Note. ω_t = Omega total.

Internal consistency estimates ranged from good to excellent (.73-.88) for collisions and speed, indicating that both normal and fog conditions captured stable performance-related behaviours. Speed was used as a proxy for time, justified by its strong negative correlation with actual time during both the normal (r = -0.69, p < .001) and fog (r = -0.60, p < .001) conditions.

2.3.1.2 Communication Measures

The descriptive statistics and reliability estimates for the coded communication behaviours are presented in Table 2.3, and the corresponding statistics for the volume-based communication variables are shown in Table 2.4. These variables were calculated on a subset of 53 dyads. We could not compute them for 27 dyads because at least one member's recording of communication was inaudible or missing. All subsequent analyses involving communication variables were conducted on this 53-dyad subset.

Table 2.3. *Descriptive Statistics and Internal Consistency Estimates for the Communication Behaviours (N=53 Dyads)*

	Normal			Fog			
	ω_t	Mean	SD	ω_t	Mean	SD	*t*
Driver							
Observation or command instruction	.90	0.25	0.20	.81	0.34	0.26	-5.13***
Inquiry	.80	0.19	0.13	.83	0.24	0.15	-4.43***
Frustration	.89	0.13	0.14	.90	0.16	0.17	-2.91**
Navigator							
Helpful observation	.86	0.41	0.23	.81	0.48	0.25	-3.39**
Harmful observation	.71	0.02	0.03	.54	0.03	0.03	-2.14*
Helpful command instruction	.90	0.48	0.32	.90	0.55	0.40	-2.63*
Harmful command instruction	.43	0.04	0.04	.78	0.06	0.06	-1.74
Inquiry	.77	0.07	0.07	.81	0.10	0.10	-2.90**
Redundant	.76	0.10	0.08	.75	0.10	0.08	-0.25

Note. ω_h = Omega total.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Importantly, there were no significant differences between the full sample of dyads and the subset of dyads on the simulation-derived performance measures or individual difference variables. The software failure was identified when coding was completed, and further data collection was not possible. Still, the remaining sample size is consistent with the existing literature on collective performance which range from 15 to 43 dyads (Bahrami et al., 2010; Gorman et al., 2005; Gorman et al., 2006; Glynn & Henning, 2000; Koriat, 2015; Sniezek & Henry, 1989).

Because the distance travelled during the normal (Mean = 7113, SD = 1397) and fog conditions (Mean = 4992, SD = 1080) differed significantly ($t_{213} = 30.75$, $p < .001$), and because longer distances offered more opportunities for communication, we computed ratios for each communication variable (i.e., frequency per 100 units of distance). All descriptive statistics and t-tests presented in Table 2.3 and 2.4 were based on these distance-adjusted ratios. On average, the driving simulation took 27.76 minutes (SD = 7.29) to complete and approximately the same amount of time was spent in the normal (Mean = 14.33 minutes, SD = 3.63) and fog conditions (Mean = 12.44 minutes, SD = 4.06). As expected, most communication behaviours occurred more frequently during fog conditions. The exceptions were harmful command instruction and redundant which did not differ between the two conditions. Internal consistency estimates were generally acceptable (> .60), except for harmful observation during the fog condition ($\omega_t = .54$) and harmful command instruction during the normal condition ($\omega_t = .43$).

Table 2.4. Descriptive Statistics and Internal Consistency Estimates for Volume-Based Communication (N=53 Dyads)

	Normal			Fog			
	ω_t	Mean	SD	ω_t	Mean	SD	*t*
Duration (seconds)							
Team	.85	13.10	9.58	.70	7.83	5.42	6.95***
Driver	.87	5.31	5.62	.79	3.06	2.67	4.59***
Navigator	.73	7.77	5.79	.64	4.76	3.35	5.37***
Talking turns							
Team	.87	1.57	0.76	.85	1.90	0.91	-5.20***
Driver	.86	0.45	0.28	.80	0.58	0.34	-5.52***
Navigator	.87	1.13	0.55	.87	1.32	0.64	-4.01***

Note. ω_t = Omega total.
***$p < .001$

For the volume of communication measures, there were two consistent patterns: 1) there was a greater duration of speech in the normal than the fog condition; and 2) a greater

number of talking turns in the fog compared with the normal condition. Internal consistency estimates were acceptable ($\omega_t > .60$) for all volume of communication variables.

2.3.1.3 Individual Differences Measures

Descriptive statistics and internal consistency estimates for all individual difference variables are presented in the Tables A1 and A2 in Appendix A to maintain the focus on the performance and communication metrics. The means and standard deviations were consistent with previous studies involving Australian undergraduate students (Blanchard et al., 2020; Jackson et al., 2016a, 2016b; Jackson et al., 2017). All reliability estimates were acceptable for research purposes except repeat errors (team $\omega_t = .57$ and driver $\omega_t = .29$), switch errors (all levels, $\omega_t = .28$ - $.54$), congruent errors (all levels, $\omega_t = .39$ - $.54$), and incongruent errors (team $\omega_t = .55$ and driver $\omega_t = .44$). As these inhibitory control and cognitive flexibility variables demonstrated poor internal consistency they were removed from subsequent analyses. We had two different metrics for each of these executive function constructs: errors and response times. The response time measures demonstrated excellent internal consistency (ranging from $\omega_t = .88$ - $.95$) and were retained for our analyses. Internal consistency estimates for the personality measures ranged between $\omega_t = .47$ - $.79$ for individuals and $\omega_t = .58$ - $.79$ for dyads. Some of the internal consistency estimates for individuals were low, however, only dyadic measures were used as covariates in regression models.

2.3.2 Exploratory Factor Analysis using Communication Variables

To identify patterns of dyadic communication, we conducted EFA separately for each operational condition (normal and fog). This allowed us to extract latent factors representing common communication behaviours across laps. Tables 2.5 and Table 2.6 present correlation matrices and factor loadings for the communication variables under normal and fog conditions, respectively.

Table 2.5. *Communication Intercorrelations and EFA Results for the Normal Condition*

	Pearson *r* correlations								Factor loadings		
Communication	2	3	4	5	6	7	8	9	1	2	h^2
Inquiry (navigator)	.29*	.73***	.36**	.49***	.35**	.01	.14	.39**	**.89**	-.19	.67
Inquiry (driver)		.41**	.62***	.35**	.26	.21	.25	.10	**.64**	.04	.43
Observation or command instruction (driver)			.53***	.43**	.39**	.12	.29*	.23	**.89**	-.09	.72
Helpful observation (navigator)				.38**	.16	.29*	.20	.14	**.71**	.00	.50
Helpful command instruction (navigator)					.36**	.01	.33*	.36**	**.72**	.00	.52
Harmful command instruction (navigator)						.48***	.29*	.27	.22	**.57**	.49
Harmful observation (navigator)							.50***	.35*	-27	**.99**	.82
Frustration (driver)								.26	.00	**.74**	.55
Redundant (navigator)									.18	**.47**	.33

Note. Factor loadings >.30 are in bold. h^2= communality.
***$p < .001$, **$p < .01$, *$p < .05$

A pattern of small to large positive correlations was observed between most variables in the normal and fog conditions. We conducted a Principal Component Analysis (with Promax rotation) on the communication measures during both conditions. We used Principal Component Analysis instead of Principal Axis Factoring because an ultra-Heywood case was detected using factor analysis. Parallel analysis suggested a two-factor solution that explained 55.89% of the common variance for the normal condition and 51.76% of the common variance for the fog condition.

Consistent with the expectations, for the normal condition, all the helpful and inquiring communication variables loaded positively on the first factor and all harmful and redundant communication variables loaded positively on the second factor. These two factors were named Helpful Exchange and Harmful Navigator, respectively, and had a positive, moderate strength correlation ($r = .42$, $p < .01$), indicating that dyads tended to engage in both helpful and harmful exchanges.

Table 2.6. *Communication Intercorrelations and EFA Results for the Fog Condition*

	Pearson *r* correlations								Factor loadings		
Communication	2	3	4	5	6	7	8	9	1	2	h^2
Inquiry (navigator)	.23	.64***	.34*	.42**	.45***	.03	.11	-.10	**.95**	**-.36**	.73
Inquiry (driver)		.35*	.58***	.33*	.09	.16	.29*	.29*	**.32**	**.47**	.46
Observation or command instruction (driver)			.47***	.44***	.29*	.19	0.22	.10	**.80**	-.03	.63
Helpful observation (navigator)				.38**	.13	.28*	0.06	.04	**.59**	.14	.44
Helpful command instruction (navigator)					.62***	.05	0.27	.10	**.76**	.02	.58
Harmful command instruction (navigator)						.25	.28*	.16	**.59**	.10	.40
Harmful observation (navigator)							0.20	.25	-.04	**.61**	.36
Frustration (driver)								.31*	.03	**.64**	.42
Redundant (navigator)									**-.31**	**.88**	.64

Note. Factor loadings >.30 are in bold. h^2= communality.
***$p < .001$, **$p < .01$, *$p < .05$

For the fog condition, the pattern of loadings was broadly similar to the normal condition, but with some subtle differences. Helpful exchange was defined by helpful and inquiring behaviours which loaded positively. Notably, the navigator's harmful command instruction also loaded positively, while redundant communication loaded negatively on this factor (marginally > .30). All negative communication variables except harmful command instruction loaded positively on the Harmful Navigator factor, as well as driver's inquiry which loaded positively and navigator's inquiry which loaded negatively. These two factors had a positive, moderate correlation with each other ($r = .43$, $p < .01$).

The extracted factors were highly correlated between the normal and fog conditions (Helpful Exchange, $r = .83$, $p < .001$; and Harmful Navigator, $r = .78$, $p < .001$). These large correlations suggest that the underlying structure of communication behaviours was stable across conditions, despite the subtle differences in the loadings. The extracted communication factor scores were used in all subsequent analyses as predictors of performance.

2.3.3 Performance During the Simulation

2.3.3.1 Individuals vs. Dyads

To test hypotheses 1a-1d, we conducted two-way mixed design ANOVAs examining the effects of grouping (between subjects: individuals vs. dyads) and operational condition (within-subjects: normal vs. fog) on the two-performance metrics: collisions and speed. Given unequal sample sizes (individual vs dyads), we used Type II Sums of Squares (Langsrud, 2003) for all ANOVA models.

Participants travelled significantly farther during the normal condition (Mean = 7113, SD = 1397) than during the fog condition (Mean = 4992, SD = 1080; (t_{213} = 30.75, p < .001), and distance was significantly correlated with collisions during both the normal (r = .36, p < .001) and fog (r = .44, p < .001) conditions. To address this, we computed a collision ratio to represent the average number of collisions per unit of distance travelled. There was no significant relationship between distance travelled and speed during the normal (r = -.06, p = .36) or fog conditions (r = .01, p = .93) so we used the original speed variables in these analyses. We used collisions ratio in all subsequent analyses which represented the number of collisions per 100 units of distance. Figure 2.6 displays mean values for collisions ratio and speed for individuals and dyads. Table 2.7 reports the corresponding AONVA results.

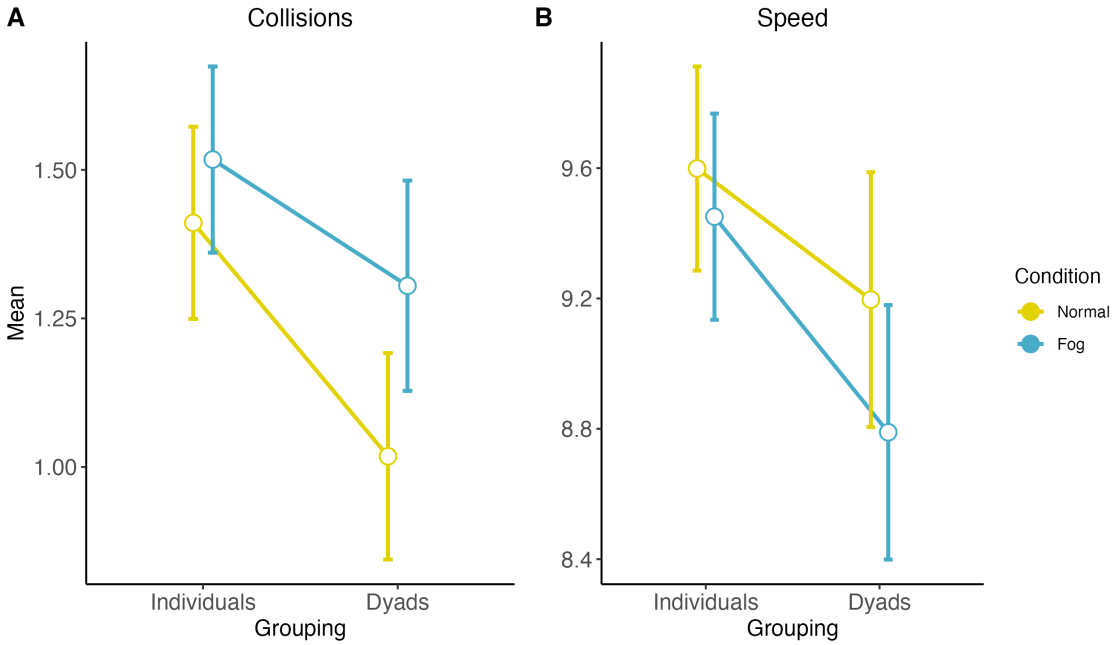
Dyads had significantly fewer collisions than individuals overall and significantly more collisions occurred during the fog than the normal condition. The interaction effect was also significant. Supporting hypotheses 1a and 1c, collisions were significantly lower for dyads than individuals during the normal condition (t_{216} = -3.10, p < .01) and there was no significant difference between individuals and dyads during the fog condition (t_{216} = -1.70, p = .09). Supporting hypotheses 1b and 1d, speed was significantly lower for dyads than individuals and significantly lower for the fog compared with the normal condition. The interaction effect was not significant.

Table 2.7. *Results of ANOVAs for Collisions and Speed*

	Individuals Mean (SD)	Dyads Mean (SD)	Mean difference	F (1,212)	η^2
Collisions ratio					
Grouping	1.46 (.94)	1.16 (.81)	0.30	6.43*	.027
Operational condition	1.26 (.92)	1.44 (.89)	-0.17	20.46***	.010
Interaction	-	-	-	5.16*	.002
Speed					
Grouping	9.52 (1.86)	8.99 (1.79)	0.53	4.63*	.020
Operational condition	9.45 (1.83)	9.20 (1.86)	0.25	11.11**	.004
Interaction	-	-	-	2.92	-

***$p < .001$; **$p < .01$; *$p < .05$

Figure 2.6. *Mean Collisions (A) and Speed (B) for Individuals and Dyads During Both Conditions*



2.3.3.2 Communication as a Predictor

To examine Aim 2 and test hypotheses 2a-2d, a series of hierarchical regression analyses were conducted to evaluate whether the extracted communication factors predicted dyadic performance, after accounting for covariates. Separate hierarchical regression models were estimated for each operational condition (normal and fog) and each dependent variable (collisions and speed).

To retain statistical power while accounting for multiple predictors, we reduced the covariates down to a smaller number of components using EFA. These extracted components were: Executive Function Time which was composed of the response time variables for repeat time, switch time, congruent time, and incongruent time; and Competence which was composed of fluid intelligence, cognitive confidence, and working memory accuracy (see Tables A3 and A4 in Appendix A for details). In addition, the proportion of females in each dyad, Neuroticism, and simulation confidence were included as control variables. Covariates were entered in blocks 1 and 2, followed by the two communication factors, Helpful Exchange and Harmful Navigator, in Block 3. The results of the analyses are presented in Table 2.8[2].

In the normal condition, collisions were significantly predicted by Harmful Navigator ($\beta = .43$, $p < .01$), while Helpful Exchange was not a significant predictor. The addition of the communication factors significantly improved model fit ($\Delta R^2 = .17$, $p < .01$). These results support Hypothesis 2c but not Hypothesis 2a. In the fog condition, Harmful Navigator had a marginal association with collisions ($\beta = .29$, $p = .07$) but did not reach statistical significance. Helpful Exchange was not a significant predictor. Thus, partial support was found for Hypothesis 2c, while Hypothesis 2a was not supported.

For speed, the communication factors did not significantly predict performance in the normal condition, although Helpful Exchange approached significance ($\beta = –.28$, $p = .06$). In contrast, during the fog condition, Helpful Exchange significantly predicted lower speed ($\beta = –.41$, $p < .01$) and the overall model improvement was statistically significant ($\Delta R^2 = .13$, $p < .05$). This provides support for Hypothesis 2b. Harmful Navigator did not predict speed in either condition, therefore Hypothesis 2d was not supported.

[2] We ran the same analyses including Agreeableness, Conscientiousness, Extraversion, and Openness to Experience. None of these additional personality traits were significant predictors of the performance metrics and the results were largely the same as those reported in Table 2.8. There was one exception, for the normal condition Helpful Exchange was a significant predictor of Speed. We excluded these personality traits to maintain adequate power in the analyses.

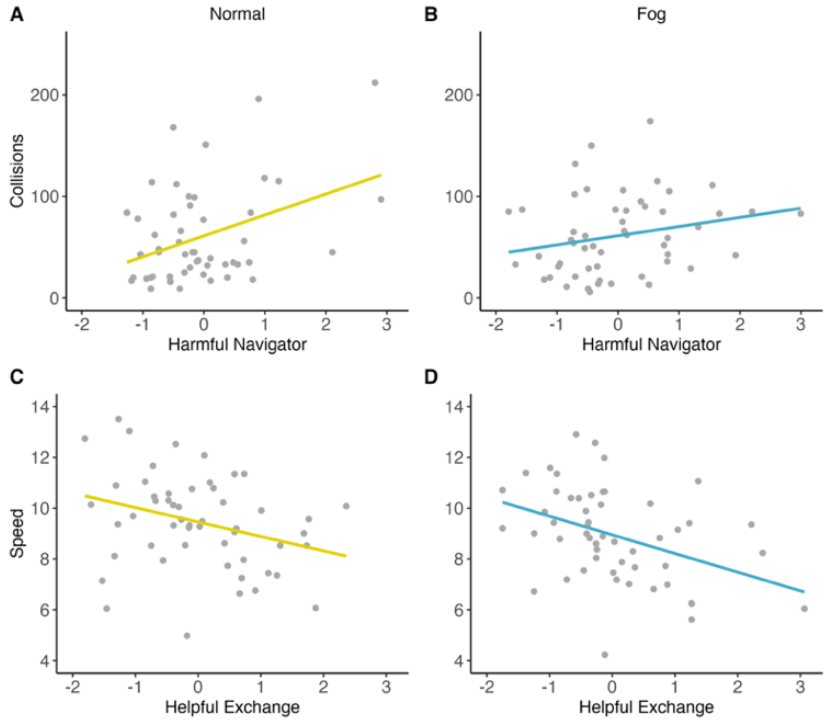
Table 2.8. *Hierarchical Regression Analyses Predicting Performance Using the Quality of Communication (N=53 Dyads)*

		Collisions								Speed							
		Normal				Fog				Normal				Fog			
Block	Predictor	R	R^2	ΔR^2	β	R	R^2	ΔR^2	β	R	R^2	ΔR^2	β	R	R^2	ΔR^2	β
1		.30	.09	.09*		.23	.05	.05†		.43	.19	.19**		.42	.17	.17**	
	Proportion Females				0.30*				0.24†				-0.43**				-0.42**
2		.33	.11	.02		.39	.15	.09		.52	.27	.08		.46	.21	.04	
	EF Time Factor				0.07				-0.02				0.20				0.02
	Competence Factor				0.01				-0.04				-0.05				-0.04
	Neuroticism				0.11				0.31*				0.18				0.19
	Simulation Confidence				0.05				0.03				0.10				0.02
3		.53	.28	.17**		.46	.22	.07		.58	.33	.06		.59	.34	.13*	
	Helpful Exchange				0.01				-0.03				-0.28†				-0.41**
	Harmful Navigator				0.43**				0.29†				0.08				0.19

Note. β = standardized regression coefficient.
** $p < .01$, * $p < .05$, † $p < .10$.

Scatterplots illustrating the relationships between the communication factors and each performance metric in both conditions are presented in Figure 2.7.

Figure 2.7. *Scatterplots for Harmful Navigator and Collisions (A and B) and Helpful Exchange and Speed (C and D) During Both Conditions*



2.3.3.3 Volume of Communication

To address aim 3 and test hypotheses 3a and 3b, we conducted a second set of hierarchical regression analyses using traditional volume-based measures of communication (i.e., duration of speech and number of speaking turns) instead of the extracted communication factors. The purpose was to qualitatively compare whether our novel communication quality metrics were stronger predictors of accuracy, while volume-based metrics were stronger predictors of speed.

As with previous models, block 1 included the proportion of females, followed by Executive Function time, Competence, and Neuroticism in block 2. The volume of communication variables were entered in block 3. Due to strong multicollinearity between duration and number of speaking turns for normal ($r = .72$, $p < .001$) and fog ($r = .69$,

$p < .001$) condition, each volume variable was entered into separate regression models. The results of block 3 are presented in Table 2.9.

Table 2.9. *Hierarchical Regression Analyses Predicting Performance Using Volume of Communication Measures (N=53 Dyads)*

	Normal				Fog			
Predictor	R	R^2	ΔR^2	β	R	R^2	ΔR^2	β
Collisions								
Duration of communication	.33	.11	.00	0.05	.39	.15	.00	0.04
Number of talking turns	.41	.17	$.06^{\dagger}$	$0.25^{\dagger}$	.40	.16	.02	0.12
Speed								
Duration of communication	.63	.40	$.14^{**}$	-0.40^{**}	.63	.40	$.19^{***}$	-0.48^{***}
Number of talking turns	.58	.33	$.07^{*}$	-0.27^{*}	.56	.32	$.10^{*}$	-0.33^{*}

Note. β = standardized regression coefficient.
*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$

The regression analyses revealed that neither of the volume of communication variables accounted for a significant amount of variance in collisions, after accounting for covariates. Duration of speech was not a significant predictor of collisions in either condition, while the number of speaking turns approach significance in the normal condition (β = .25, p = .08) but not in the fog condition. These results suggest that communication volume has limited utility in predicting dyadic accuracy outcomes. In contrast, volume-based predictors were significantly associated with speed. Duration of speech was a significant negative predictor of speed in both the normal (β = –.40, $p < .01$) and fog (β = –.48, $p < .001$) conditions. Likewise, number of speaking turns significantly predicted slower speed in both conditions, although to a lesser extent.

A qualitative comparison of the regression coefficients from the two sets of analyses supports Hypotheses 3a and 3b. The Harmful Navigator factor was a stronger predictor of collisions in the normal condition (β = .43) than either duration (β = .05) or speaking turns (β = .25). Conversely, communication duration was a stronger predictor of speed (β = –.48 in fog) than the Helpful Exchange factor (β = –.41). Thus, communication quality was more

strongly related to accuracy, while communication volume was marginally more strongly related to speed.

2.4 Discussion

This study was the first to examine the 2HBT1 effect in a dynamic driving simulation using naïve, asymmetrical, and distributed dyads under varying operational conditions. It also introduced a novel method for quantifying the quality of communication and demonstrated that this measure was a stronger predictor of dyadic accuracy than traditional volume-based metrics. These findings provide important insights into when and how dyads outperform individuals in dynamic environments and highlight the utility of assessing the quality of communication.

2.4.1 Two Heads are Not Always Better Than One

As expected, the 2HBT1 effect depended on characteristics of the task (i.e., operational condition) and the performance metric (i.e., accuracy vs. speed).

For accuracy, dyads had fewer collisions than individuals during the normal condition, but this advantage disappeared during the fog condition. These findings suggest that under familiar and stable conditions, where team members share access to similar task-relevant information, dyads perform better. However, when conditions change unexpectedly and cognitive demands increase, dyads may revert to individual level performance. This reduction in the two-heads advantage is a failure to capitalise on the navigator's informational advantage. The sudden onset of fog disrupted the driver's visibility and likely impaired the dyad's shared understanding of the task environment. Despite the navigator's broader access to environmental information, dyads did not adapt in ways that improved collective accuracy. Beneficial adaptation would involve increased communication and greater reliance upon the navigator for timely and helpful updates about the driver's

surroundings. However, we observed that both drivers and navigators spoke for less time during the fog condition compared to the normal condition. This is likely because the fog increased the driver's cognitive load, forcing them to divert cognitive resources away from communication and toward monitoring the environment and operating the vehicle.

Prior research (e.g., Cooke et al., 2009; Gorman et al., 2005) has shown that when operational conditions change, groups must re-establish shared situation awareness to maintain performance. Our findings suggest that this adaptation did not occur within the short duration of the fog condition, perhaps because drivers responded more like individuals than members of a coordinated team. This may be due to the naïve or ad hoc nature of our dyads, the brevity of the fog event, or the absence of training. Unexpected events in real-world tasks vary in duration, some last only minutes like our task (e.g., unexpected weather conditions or equipment malfunction), while others may persist for weeks or months (e.g., natural disasters). We suspect that longer disruptions may provide dyads with sufficient time to adapt and recover the 2HBT1 effect. Understanding how dyads respond across a range of event durations is critical. It is possible that our brief fog event did not allow enough time for dyads to adapt and regain their performance advantage. Future research should test this hypothesis using unexpected events of varying lengths.

For speed, dyads drove more slowly than individuals in the fog condition, but not in the normal condition. Interestingly, a speed-accuracy trade-off did not account this pattern. Compared to the normal condition, dyads in the fog both reduced their speed and increased their collisions, while individuals did not differ between conditions. The slower speed observed in dyads may reflect a compensatory strategy in response to the increased cognitive demands of the fog condition. This aligns with prior findings that communication imposes a cognitive cost (MacMillan et al., 2004). Furthermore, prior research on distracted driving (e.g., Haigney et al., 2000) shows that individuals tend to reduce their speed when engaging

in task-irrelevant communication via mobile phones. Our results suggest that similar compensatory mechanisms occur in dyads engaged in task-relevant communication but only when the task environment changes unexpectedly. In the fog condition, drivers may have reduced their speed to maintain safety while processing incoming communication from the navigator.

Taken together, our findings highlight the importance of context when evaluating the benefits of teamwork. The 2HBT1 effect emerged for accuracy under stable conditions but not when the task conditions changed and became more cognitively demanding. For speed, dyads were consistently slower than individuals. Whether two heads are better, the same, or worse than one depends on the operational conditions and the performance metric most relevant to the task's goals. If accuracy is the priority and the environment is stable, dyads may be beneficial. However, in volatile environments, individuals may be a better choice because they achieve comparable performance while using fewer resources. When speed (or time) is the key performance metric, individuals are likely more effective in both stable and unstable conditions. Overall, the use of asymmetrical and distributed dyads in dynamic tasks should be carefully considered, as two naïve heads are not often better than one.

2.4.2 Communication and Dyadic Performance

Our second novel contribution was to investigate the relationship between communication and dyadic performance using both novel communication quality and traditional volume-based metrics. The quality of communication, particularly harmful communication, was a predictor of accuracy in the normal condition. Dyads that engaged in more ill-timed or inaccurate speech experienced more collisions, whereas helpful communication was not associated with accuracy. This asymmetry may reflect a negativity bias, where the cost of poor communication outweighs the benefits of good communication. It may also be due to the prevalence of common (rather than unique) knowledge during the

normal condition. Since both members had access to similar information, communication likely focused on knowledge both members already knew. Prior research shows that groups tend to spend more time discussing shared knowledge, which has less impact on group outcomes than unique knowledge (Stasser & Titus, 1985; Mesmer-Magnus & DeChurch, 2009). In this context, harmful communication, which was incorrect and/or ill-timed, may have appeared to drivers as unique or important, thus capturing their attention and influencing behaviour. Interestingly, no communication variables predicted accuracy in the fog condition, despite the navigator's informational advantage. This may be because dyads failed to adapt to the changed conditions and drivers operated more like individuals. If drivers disregarded the navigator's input when under pressure, then, as we observed, communication would not impact accuracy.

As expected, communication quality predicted speed in the fog condition. Dyads that engaged in more helpful communication drove more slowly, likely as a compensatory strategy to manage the increased cognitive demands of processing that information. This effect did not emerge in the normal condition, possibly because drivers had more cognitive capacity available to manage driving and communication simultaneously. Harmful communication was not associated with speed, likely due to its lower frequency.

2.4.3 Comparing Communication Metrics

Next, we conducted a qualitative comparison to assess which type of communication metric best predicted accuracy and speed. As expected, the quality of communication was more important for accuracy, whereas the volume of communication was more important for speed. Specifically, poor communication was the strongest predictor of collisions, and the duration of speech was the strongest predictor of speed. These findings reinforce the value of assessing the content of communication, not just its frequency. Dyads that speak more may not necessarily perform better, particularly if the communication is irrelevant, inaccurate, or

poorly timed. If anything, communication may always carry a performance cost to speed when asymmetrical and distributed dyads work together on a dynamic task.

Our novel coding approach and simulation design allowed use to map communication onto conditions and laps enabling a psychometric evaluation of different communication measures. These analyses revealed stable communication structures across both conditions. Our methodology could be adopted by future research to more objectively quantify communication quality, or even volume-based metrics, which are rarely subjected to psychometric assessment.

2.4.4 Covariates and Dyadic Performance

There was no systematic pattern between the theory-driven control variables (executive function, intelligence, cognitive confidence, simulation confidence, and personality) and performance metrics. However, one pattern emerged: Dyads with a higher proportion of females had more collisions and drove at lower speeds than dyads with a lower proportion of females.

These findings are consistent with prior research on driving. For example, females tend to have more accidents while manoeuvring a vehicle through traffic (Laapotti & Keskinen, 2004) and tend to drive more cautiously (Sarma et al., 2013) compared with males. These findings for accuracy, however, conflicts with recent research on collective intelligence, which suggests that dyads with a higher proportion of females tend to perform better than those with a lower proportion of females across a broad range of tasks. However, this previous research critically differs from our study, as it used symmetrical dyads and did not include a dynamic task in their test battery (Woolley et al., 2010).

2.4.5 Limitations and Future Directions

Several limitations of our study should be acknowledged. First, our sample size limited the ability to extract communication factors for each lap. Since communication is a dynamic process, this restricted our ability to examine how patterns of communication quality evolve over time. The small sample size also prevented us from testing whether personality traits served as mediators or moderators of the relationships between communication and performance. Future research should recruit larger samples to explore these possibilities, as personality may have important indirect effects on dyadic outcomes.

Second, our dyads were composed of participants with no prior experience working together or using the driving simulation. While this reflects common real-world scenarios involving naïve dyads, the relationships we observed between communication and performance may differ for dyads composed of task experts, or well-established dyads. These effects might also vary across tasks with lower cognitive demands. Future research should examine how these dynamics influence performance by using expert dyads and varying task complexity.

Third, our measure of simulation confidence relied on a single item administered post-task. Single items measures prevent the estimation of internal consistency and are more vulnerable to measurement error, which can reduce construct validity. Furthermore, dynamic tasks like our simulation involve thousands of small decisions, making it unlikely that a single, global confidence rating can accurately reflect participants' experiences. Future research should employ multiple in-tasks confidence probes, or at a minimum, a multi-item post-task measure designed to capture the range of experiences participants encounter during the simulation.

Fourth, while our communication quality metrics captured navigator performance, our dyadic performance metrics (outcomes) focused exclusively on the driver. Future research should incorporate other measures of navigator effectiveness, such as the proportion of time the driver's vehicle or route was visible to them.

Finally, although participants were provided with an overview of their roles and task goals, they did not receive scenario-specific training. In many real-world settings, especially high stakes environments, dyad members are trained prior to engaging with the task. Future studies should explore how training impacts the 2HBT1 effect and its relationship with communication quality for asymmetrical, distributed dyads.

Beyond these limitations, our findings illuminate several areas for future research. First, studies should test whether dyads recover the 2HBT1 effect when unexpected events last longer and allow more time for adaptation. Second, researchers could introduce a broader range of event types to examine whether our effects generalise. Third, future work could explore conditions that systematically lead to the "the two heads are worse than one" effect, as observed in static tasks (Koriat 2012, 2015). Finally, our findings should be extended to other types of dynamic tasks to determine the generalisability of our findings for performance and communication across domains.

2.4.6 Implications

Our findings for asymmetrical and distributed dyads have several implications. In our study, the 2HBT1 effect only emerged for accuracy under stable and familiar conditions, which are characterised by low uncertainty and shared access to environmental information. Under volatile conditions, like our sudden onset of fog, this advantage disappeared. For speed, dyads consistently underperformed relative to individuals. These results suggest that individuals may be more efficient than dyads under dynamic task conditions similar to our

simulation, especially when the operating environment is volatile or unpredictable. There's one exception, if accuracy is the primary performance metric and the environment is relatively stable, dyads may offer an advantage and be the better choice.

Our communication results highlight the importance of content. Communication quality was more strongly associated with accuracy, while communication volume was more strongly associated with speed. This suggests that the cognitive costs of communication should not be ignored, particularly for dyads operating in high cognitive load environments. Training dyads to communicate more efficiently may help preserve accuracy while reducing the performance cost to speed.

Furthermore, our novel measure of communication quality allowed us to evaluate its psychometric properties and directly map communication to on-task behaviour. This represents a valuable contribution to the measurement of dyadic communication, which has often relied on post-task self-reports. Future research adopting similar more objective methods may further clarify the mechanisms through which communication supports or hinders dyadic performance across varying task conditions and dyad structures.

2.4.7 Conclusion

By embedding different operational conditions within a dynamic task, this study provides new insights into the performance of asymmetrical, distributed dyads and the influence of communication. We demonstrated that dyads outperformed individuals under stable conditions but may lose this advantage when task conditions shift unexpectedly. We also showed that communication quality, particularly harmful communication, plays a critical role in determining dyadic accuracy, while communication volume more strongly predicts speed. Importantly, we developed a novel method for quantifying communication quality that

allowed us to assess its psychometric properties and examine its relationship with dyadic performance in real-time.

Together, our findings demonstrate that two heads are not always better than one. The two-heads advantage depends on task conditions, performance goals, and the nature of communication between dyad members. The use of asymmetrical and distributed dyads in dynamic task environments should therefore be carefully considered.

Chapter 3: Study 2

A Recipe for Dyadic Collective Intelligence for Well-Structured Tasks: Mix Equal Parts Cognitive Ability and Confidence Plus a Pinch of Social Sensitivity

The original manuscript for the study described in this chapter has been published in the journal *Cognitive Research: Principles and Implications*: Blanchard, M. D., Aidman, E., Stankov, L., & Kleitman, S. (2024). A Recipe for Dyadic Collective Intelligence for Well-Structured Tasks: Mix Equal Parts Cognitive Ability and Confidence Plus a Pinch of Social Sensitivity. *Cognitive Research: Principles and Implications*, 10, 63. https://doi.org/10.1186/s41235-025-00655-0.

3.1 Introduction

Collective decision making is a complex process influenced by human interaction, cognitive ability, and metacognitive confidence (e.g., Bahrami et al., 2010; Hill, 1982; Hinsz, 1990; Zarnoth & Sniezek, 1997). Woolley et al. (2010) introduced the concept of *collective intelligence* as a group's ability to perform across a wide variety of tasks, emphasizing nuanced factors like one's ability to perceive others' emotions and interaction processes, rather than cognitive ability. This seminal work provided a new perspective for understanding why some groups thrive when making collective decisions and others do not. However, subsequent research (e.g., Graf-Drasch et al., 2022; Rowe et al., 2021) suggests Woolley et al.'s findings may be task dependent and limited in generalisability, particularly for well-structured tasks which have a single, objective solution.

To address these limitations, our study re-examined collective intelligence by focusing on dyads and employing well-structured tasks guided by the Cattell-Horn-Carroll (CHC) model of intelligence. This approach allowed for a more precise assessment of the relationship between individual intelligence and dyadic collective intelligence. Additionally, we explored the role of metacognitive confidence, given its central influence on group

decision-making outcomes (e.g., Bahrami et al., 2010; Kerr & Tindale, 2004; Koriat, 2015), and how it may uniquely contribute to collective intelligence.

Our primary aim was to clarify the relationship between individual intelligence and collective intelligence for dyads on well-structured tasks. We hypothesized that individual intelligence would strongly predict dyadic collective intelligence, challenging previous findings that social factors are more critical than intelligence. Furthermore, we investigated the relationship between metacognitive confidence and collective intelligence to understand how individuals' confidence levels impact collective performance. Finally, we applied Latent Profile Analysis (LPA) as a novel, person-centred approach to identify distinct psychological profiles of dyads based on their individual and dyadic scores on intelligence, metacognitive confidence, and bias (the accuracy of confidence). This methodology allowed us to differentiate dyads by their performance patterns, revealing common changes in outcomes from individual to dyadic performance and offering insights into optimizing group compositions for enhanced collective intelligence.

3.1.1 Collective Intelligence

Collective intelligence broadly refers to the common finding that groups tend to perform better than individuals (e.g., Blanchard et al., 2023, Blanchard et al., 2024; Gordon, 1924; Hill, 1982; Kameda et al., 2022; Kerr & Tindale, 2004; Kurvers et al., 2016; Williams & Sternberg, 1988). Various conceptualizations of collective intelligence exist, including consensus-seeking groups and the wisdom-of-crowds. The present study focused on Woolley et al.'s (2010) definition of collective intelligence for consensus seeking groups as "the general ability of the group to perform a wide variety of tasks" (p. 687), rooted in the idea that collective intelligence parallels individual intelligence but at the group level.

Woolley et al. (2010) modelled collective intelligence on Spearman's theory of general intelligence for individuals (Spearman, 1904), using a battery of tasks selected based on the McGrath Group Task Circumplex (McGrath, 1984). This is a taxonomy categorizing tasks by the coordination processes required for completion. Woolley and colleagues selected at least one task from each category (generation, decision-making, negotiation, and execution) to provide a broad assessment of group ability. In the first study, their tasks included brainstorming possible uses for a brick, RAPM, moral reasoning, negotiating plans for a shopping trip, and typing a difficult text under time constraints. A factor analysis across these tasks yielded a first-order collective intelligence factor ('*c*'), analogous to Spearman's '*g*' for individual general intelligence, suggesting that collective intelligence could predict group performance across varied contexts.

3.1.1.1 Original Research Findings.

In two studies involving 192 groups of three to five members, Woolley et al. (2010) found that collective intelligence accounted for approximately 43-44% of the variance in group performance across tasks. Collective intelligence also predicted performance on a criterion task (study 1: computerised checkers, r = .52; study 2: architectural design, r = .28). Their results linked higher collective intelligence scores with greater social sensitivity, equality of turn-taking (i.e., lower variance in speaking turns), and a higher proportion of female group members, although this last finding was fully mediated by social sensitivity. These findings have been replicated several times by Woolley and colleagues (Aggarwal et al., 2019; Engel et al., 2014, 2015; Kim et al., 2017; Riedl et al., 2021), suggesting that collective intelligence is distinct from individual intelligence, with a weak correlation between average individual intelligence and collective intelligence (r = .15).

These studies make two major claims. First, social factors, such as accurately perceiving others' emotions (social sensitivity) and sharing conversational turns equally, are

more important for group performance than individual intelligence. However, the finding for social sensitivity requires clarification, as the instrument used for its measurement, Reading the Mind in the Eyes test, has limitations. Research suggests it often produces a ceiling effect in neurotypical populations so it may not be effective at assessing individual differences (Black, 2017), often has low internal consistency (e.g., Harkness et al., 2010; Ragsdale & Foley, 2011), and is a mixed measure of emotion perception, vocabulary knowledge, cognitive empathy, and affective empathy (Kittel et al., 2021; Olderbak et al., 2015). This finding requires replication with a different tool that assesses emotion perception. The second major claim is that groups possess general abilities and characteristics that allow them to perform well across a wide range of contexts. This challenges the long-held belief that groups require specialised abilities and skillsets (e.g., expertise) to perform effectively in specific contexts (Cohen & Bailey, 1997; Hollingshead & Poole, 2012; Steiner, 1972).

Woolley et al.'s studies present collective intelligence as a robust construct, independent of individual intelligence, that predicts group performance across various contexts. However, the claim that collective intelligence "has been well established in the literature" (Askay et al., 2019, p. 492) is contradicted by mixed results from independent research.

3.1.1.2 Independent Research and Meta-Analytic Findings

Credé and Howardson (2016) reanalysed the data from 6 studies, finding that collective intelligence accounted for limited variance in group performance, and the collective intelligence tasks had low internal consistency. Barlow and Dennis (2016) failed to replicate a dominant collective intelligence factor for virtual, text-based groups, and social sensitivity showed inconsistent relationships with performance on each of the group tasks. Bates and Gupta (2017) observed that individual intelligence explained much of collective intelligence's variance, while Woolley's key predictors (social sensitivity, equality of turn

taking, and proportion of women) showed no significant relationship with collective intelligence. Guided by Horn and Cattell's (1966) Gf-Gc theory of individual intelligence, Rowe et al. (2024) used group tasks designed to assess individual intelligence and found that two factors, rather than one, best accounted for variance in collective intelligence. Neither social sensitivity nor equality of turn taking was related to the collective intelligence factors.

A meta-analysis by Graf et al. (2019) proposed a 3-factor structure for collective intelligence (idea generation, conflict resolution, and task execution) which challenged Woolley's one-factor model and aligns with broader group research. For example, LePine et al. (2008), in a meta-analysis of 138 studies, found a second-order model with three broad processes (transition, action, and interpersonal) and a higher-order Teamwork Process factor that was correlated with group performance. In a review, Hackman and Morris (1975) also showed that group interaction processes strongly related to group performance. These findings suggest that Woolley's collective intelligence captures teamwork processes and dynamics rather than a cognitive construct analogous to individual intelligence.

Rowe et al. (2021) provided additional insights into collective intelligence's limitations in a meta-analysis, finding a weak correlation with average individual intelligence (r = -.05 to .34). Tasks such as moral decision making, negotiating a shopping trip, and collaboratively typing are not typical intelligence measures, potentially limiting collective intelligence's relevance as a cognitive construct. These inconsistencies raise questions about collective intelligence's robustness and its dependence on the types of tasks chosen.

Laughlin (2011) distinguished between intellective and judgmental tasks, illustrating how task structure influences group decision-making. Building on Laughlin's work, Graf-Drasch et al. (2022) categorised tasks used in prior collective intelligence research as either well-structured or ill-structured and examined collective intelligence for each type. Woolley's collective intelligence emerged only for well-structured tasks, which have a clear strategy

leading to a single correct solution, unlike ill-structured tasks that are ambiguous, allow multiple solutions, and emphasise group interaction.

These findings imply that Woolley's collective intelligence may only apply to well-structured tasks, and its relationship with individual intelligence remains unclear. They also raise two critical questions: Could individual intelligence predict dyadic collective intelligence when measured using well-structured tasks? Prior research suggests this is likely, given that individual intelligence tend to relate to group performance on well-structured tasks (Bruine de Bruin, 2007, 2012, 2019; Del Missier et al., 2012). Furthermore, extensive research demonstrates the central role of individual intelligence in group performance (Barrick et al., 1998; Bell, 2007; Devine & Philips, 2001; Imbimbo et al., 2021; LePine, 2003, 2005; LePine et al., 1997; Stewart, 2006).

Second, are Woolley's key predictors (social sensitivity, equality of turn taking, and the proportion of females) related to dyadic collective intelligence for well-structured tasks? Inferring others' emotions is crucial for effective social interactions (Lopes et al., 2003), and equality of turn taking is associated with higher decision quality (Janis & Mann, 1977; Vroom & Yetton, 1973). Gender differences in communication styles, with females tending to be more supportive and males more dominant (Anderson & Leaper, 1998; Carli, 2001; Carli & Bukatko, 2000; Fishman, 1978; Leet-Pellegrini, 1980), may be more pronounced for larger groups (3-5 members) and during ill-structured tasks where more communication is generally required. Thus, for dyads, the cognitive requirements of well-structured tasks may drive gender differences. For example, there may be a small male advantage for tasks that primarily require fluid reasoning (e.g., Halpern et al., 2007), and there may be a small female advantage for tasks that primarily require verbal abilities (e.g., Reilly et al., 2019).

3.1.1.3 Methodological Considerations

Replicating collective intelligence with varied methods is crucial for assessing its robustness (Botvinik-Nezer et al., 2020). Woolley's reliance on theoretical taxonomies, like the McGrath's Group Task Circumplex, often lacks empirical support, leading to overlapping task categories that can blur construct clarity (Devine, 2002). These taxonomies often fail to distinguish between the characteristics that drive group effectiveness for different types of dyads (e.g., dyads engaged in building a house versus those engaged in conducting scientific research; Cohen & Bailey, 1997). Theoretical taxonomies typically sample items based on their subjective correspondence with perceived group characteristics (e.g., generate, choose, negotiate, and execute). This approach is guided by theory but lacks empirical validation. Bell's (2007) meta-analysis found that group task taxonomies had no moderating effect on the relationship between individual intelligence and group performance. Most studies of Woolley et al.'s (2010) collective intelligence used the McGrath Group Task Circumplex to guide the selection of group tasks. This suggests their selection of items may not adequately cover the hypothesized construct to capture the latent properties of the cognitive processes that drive group performance.

In contrast, the Cattell-Horn-Carroll (CHC) model of intelligence offers an empirically validated framework for understanding the cognitive processes involved with individual performance on a wide variety of tasks (e.g., Carroll, 1993; Horn & Cattell, 1966; McGrew, 2009). The CHC model synthesises over a century of research and encompasses 16 broad abilities, such as Fluid Reasoning (i.e., abstract reasoning that has little dependence on acquired knowledge), Crystallised intelligence (i.e., acquired knowledge that is culturally relevant), and Quantitative Knowledge (i.e., acquired knowledge about mathematics). Each of these abilities is supported by extensive psychometric evidence and has been linked to performance on a wide range of tasks.

By using the CHC model to guide the selection of well-structured tasks we can reliably assess the relationship between individual intelligence and dyadic collective intelligence. This approach allowed us to examine the cognitive abilities that are most relevant to collective intelligence. For example, Fluid Reasoning may help us understand how dyads solve novel problems, while Crystallised Intelligence can shed light on how shared knowledge and unique knowledge contributes to group decisions.

However, it's important to note that the CHC model is designed to explain cognitive abilities at the individual level and does not account for the interaction processes that emerge when people work together in groups. Interaction processes (e.g., communication patterns, coordination, and conflict resolution) play a critical role in group performance (Hackman & Morris, 1975; Mesmer-Magnus & DeChurch, 2009). These processes can lead to synergistic effects where the group's performance exceeds the sum of its parts (e.g., high collective intelligence), or conversely, to process losses where group performance is harmed (Steiner, 1972).

While the CHC model provides a strong foundation for assessing the cognitive components of dyadic performance, it was not designed to capture these dynamic social interactions. Incorporating measures of interaction processes could enhance our understanding of collective intelligence by revealing how cognitive abilities, interaction processes, and social dynamics jointly contribute to group outcomes. However, integrating these factors requires a more complex research design and is beyond the scope of the present study. Our focus is on isolating the impact of individual intelligence on dyadic collective intelligence using well-structured tasks guided by the CHC model.

The hierarchical structure of the CHC model, which includes both broad and narrow cognitive abilities, provides a nuanced framework for analysing the interplay between

individual intelligence and collective intelligence. This approach addresses previous methodological limitations by grounding our task selection in a validated theoretical model.

3.1.1.4 The Current Research

Addressing previous limitations in collective intelligence research, the primary aim of our study was to examine two critical questions: First, does individual intelligence predict dyadic collective intelligence when measured using well-structured tasks? We used parallel forms of a Fluid Reasoning test to assess both individual intelligence and collective intelligence, along with additional measures of Crystallized Intelligence and Quantitative Knowledge to measure collective intelligence. This approach allowed for a more precise investigation of the relationship between intelligence at the individual and collective levels. Prior research outside the collective intelligence paradigm indicates a moderate correlation between individual cognitive ability and group performance for these tasks (Del Missier et al., 2012; Bruine de Bruin, 2019).

H1: We hypothesised that individual intelligence would strongly predict dyadic collective intelligence for well-structured tasks, after accounting for the other key variables.

Second, do social sensitivity, equality of turn-taking, and proportion of females, identified by Woolley et al. (2010), predict dyadic collective intelligence for well-structured tasks? We measured collective intelligence using tasks that capture Fluid Reasoning, Crystallised Intelligence, and Quantitative Knowledge. Two of these tend to show small male performance advantages; thus, collective intelligence may be greater for male than female dyads.

H2a: We hypothesised that the proportion of females would negatively predict collective intelligence, after accounting for the other key variables.

This conflicts with Woolley et al.'s finding that social sensitivity fully mediated the positive relationship between proportion of females and collective intelligence. We expected that the cognitive abilities required for well-structured task success would have a stronger influence on the relationship between gender composition and dyadic collective intelligence than an indirect effect through social sensitivity.

While social sensitivity and equality of turn-taking are associated with high-quality decision-making (Janis & Mann, 1977; Vroom & Yetton, 1973), independent collective intelligence researchers have failed to replicate Woolley and colleagues' original results (Barlow & Dennis, 2016; Bates & Gupta, 2017; Rowe et al., 2024). Therefore, we did not expect these predictors to be significant after accounting for the other variables in the model.

H2b: We hypothesised that social sensitivity and equality of turn-taking would not predict collective intelligence, after accounting for the other key variables.

3.1.2 Confidence and Group Decision-Making

Research shows that confidence significantly influences both the processes and outcomes of group decisions. Sniezek and Henry (1989) found that groups tend to be more accurate and confident than individuals. Similarly, Kerr and Tindale (2004) observed that individuals with higher confidence often dominate discussions, exerting greater influence on the group's final decision compared to less confident members. Bahrami et al. (2010) demonstrated that the "two heads are better than one" effect occurred when group members shared their confidence accurately. Blanchard et al. (2020) extended this by showing that overconfident dyads increased in decision-making errors more than underconfident or well-calibrated dyads, highlighting the role of metacognitive bias in group decision-making. Bias, or the degree of over- or under-confidence in one's judgments, reflects one's ability to accurately monitor their own performance. These findings demonstrate that confidence is a critical factor in group interactions and decision-making outcomes, suggesting its potential

relevance to collective intelligence. This relationship remains underexplored in prior research.

Confidence is sometimes overlooked due to its close relationship with accuracy. However, confidence reflects distinct psychological processes. Koriat (2024) found that confidence monitored the likelihood of response replicability rather than accuracy. High confidence indicated that an individual would likely choose the same response if faced with the same question again, even if it was incorrect. This has implications for groups: high-confidence groups are likely to show consistency over time with similar tasks, whereas low-confidence groups may behave inconsistently. It is our belief that measuring both accuracy and confidence is essential for comprehensive group research.

In the present study, we extended previous findings by exploring the distinct roles of individual confidence and intelligence in predicting collective intelligence. Our second aim was to examine the relationship between individual confidence and dyadic collective intelligence.

H3: We hypothesised that individual confidence would positively predict collective intelligence, after accounting for other key variables.

Based on Koriat's (2024) work, our third aim was to compare the predictors of collective confidence with those of collective intelligence to demonstrate that it captures unique information about dyads beyond collective intelligence alone. We expected that the relationships between collective confidence and the predictors would differ from those between collective intelligence and the same predictors.

H4: We hypothesised that collective confidence would have a distinct pattern of relationships with individual intelligence, individual confidence, and the other key variables, compared to collective intelligence.

3.1.3 Profiling Dyadic Performance

The literature has predominantly used a variable-centred approach, focusing on average relationships between variables across the entire sample, to investigate collective intelligence. This provides valuable insights into general trends and correlations; however, it may overlook important differences between distinct types of groups. For example, Woolley at al. (2010) used an analytic approach that assumed the same relationships between variables applied equally to all groups, potentially hiding meaningful heterogeneity about how individual and group characteristics interact to influence collective intelligence.

To address this limitation and extend previous research, we adopted a person-centred approach by using LPA to identify clusters of dyads with similar psychological profiles across intelligence, confidence, and bias, measured at both the individual and collective levels. LPA allowed us to identify subgroups within the sample that shared unique behaviour patterns across these variables. This approach provides a more nuanced understanding of the different types of dyads. For example, LPA can reveal profiles of dyads that outperform their individual members (*two heads are better than one* effect) or dyads that collectively underperform compared to their members working alone (*two heads are worse than one* effect). By including both individual and dyadic variables in the LPA model, we can capture shifts between individual and collective responses, thus addressing the heterogeneity of dyads providing valuable insights into possible distinct subgroups.

This method is a strategic departure that extends the approach of Woolley et al. (2010) because it simultaneously accounts for individual differences and group dynamics. It contributes to the theoretical framework of collective intelligence by providing a more

detailed “recipe” of the ingredients that are associated with different types of dyadic performance. This information is essential for the formation of effective dyads and for developing interventions that enhance outcomes.

Our fourth aim was to identify psychological profiles of dyads using clustering based on intelligence, confidence, and bias. Given the exploratory nature of this analysis, we could not predict the exact number or nature of profiles that LPA would extract. However, we anticipated the emergence of multiple profiles, including at least one showing a *two heads are better than one* effect. Other possible profiles might reflect *two heads are the same as one effects (e.g.,* high or low intelligence at both levels) and a *two heads are worse than one* effect. Furthermore, additional profiles could be shaped by variations in confidence and bias.

Our fifth aim was to examine differences between these distinct dyadic profiles on the key variables identified by Woolley and colleagues, as well as the individual difference variables described in the following section. This aim allowed us to explore how distinct dyadic profiles related to individual differences, potentially informing the selection of dyad members to improve dyadic performance or to prevent ineffective pairings.

3.1.4 Individual Differences

Working in a group is a complex task, and individual differences can influence group processes and outcomes. Key constructs include working memory, essential for holding information in mind for mental tasks (Baddeley, 1992) and Big Five personality, as meta-analyses show that Agreeableness, Conscientiousness, and Openness to Experience positively correlate with group performance (Bell & Kozlowski, 2002; Peeters et al., 2006). In this study, we examined the relationships between the LPA profiles and these individual difference variables.

3.1.5 The Present Study

To investigate our hypotheses (presented in Table 3.1) and two exploratory aims, we employed well-structured tasks aligned with the CHC model. Individual intelligence and confidence were assessed using a Fluid Reasoning task, while collective intelligence and collective confidence were assessed using three tasks capturing Fluid Reasoning, Crystallised Intelligence, and Quantitative Knowledge. Following Woolley et al., (2010), we conducted a confirmatory factor analysis (CFA) to extract a single collective intelligence factor and a single collective confidence factor. We then fit two hierarchical regression models with the collective intelligence factor as an outcome variable in model 1 and the collective confidence factor as an outcome in model 2 and the following predictors: social sensitivity, equality of turn taking, number of females, individual intelligence, individual confidence, and the individual differences variables.

Given that individuals are nested within dyads, a multilevel model is typically recommended to account for interdependence within dyads (Gonzalez & Griffin, 2023; Kenny et al., 2006). Multilevel modelling requires variance at both the within-dyad and between-dyad levels for the outcome and predictors. In our study, however, collective intelligence was measured at the dyadic level through consensus-based responses, meaning there was no within-dyad variance in collective intelligence. Kenny & Kashy (2014; p. 591) note, “if only a single outcome is obtained for each group, then group is treated as the unit of analysis, and special analytic methods are not required.” Therefore, in our case, the dyad is treated as the unit of analysis, and nesting individuals within dyads in a multilevel model not appropriate.

We did not expect social sensitivity to have a significant positive relationship with collective intelligence. However, if Woolley et al.’s (2010) finding was replicated, we would conduct a mediation analysis to investigate the direct and indirect relationship between the

proportion of females and collective intelligence, with social sensitivity as a mediator. We expected to find a negative direct relationship between the proportion of females and collective intelligence, contrasting with Woolley and colleagues, who reported that social sensitivity fully mediated a positive relationship.

We then fit an LPA model to identify clusters of dyads with similar psychological profiles across intelligence, confidence, and bias, measured at both the individual and collective levels. Finally, we conducted a series of between-subjects ANOVAs to test for differences between the identified profiles on Woolley et al.'s (2010) key variables (social sensitivity, equality of turn taking, and proportion of females) and the individual differences measures.

Table 3.1. *The Hypotheses for Study 2*

Number	Hypothesis
H1	We hypothesised that individual intelligence would strongly predict dyadic collective intelligence for well-structured tasks, after accounting for other key variables.
H2a	We hypothesised that the proportion of females would negatively predict collective intelligence, after accounting for the other key variables.
H2b	We hypothesised that social sensitivity and equality of turn-taking would not predict collective intelligence, after accounting for the other key variables.
H3	We hypothesised that individual confidence would positively predict collective intelligence, after accounting for other key variables.
H4	We hypothesised that collective confidence would have a distinct pattern of relationships with individual intelligence, individual confidence, and the other key variables, compared to collective intelligence.

3.2 Method

3.2.1 Participants

In return for partial course credit, 244 Australian undergraduate psychology students completed the study (158 females, 86 males, mean age = 20.82, SD = 4.12). Thirty-four participants were excluded from analyses who did not complete the protocol due to Qualtrics platform or computer crashes, timing constraints, or non-genuine attempts. Non-genuine attempts were identified through lab observations (where dyads did not engage as instructed),

rapid response times suggesting random guessing, and self-reports from participants acknowledging non-genuine participation.

The final sample included 210 participants (133 females, 77 males, mean age = 20.79, SD = 4.33) who completed the study as 105 two-person groups. These dyads were formed as minimal groups. That is, participants were randomly allocated to dyads, and the majority (89%) did not know each other prior to participation in the study. However, 12 dyads indicated they had met before participation, with half meeting within the last 6 months. Given the small number of dyads that knew each other prior to the study and a non-significant correlation with collective intelligence (r = -.14, p = .15), we did not include familiarity in our analyses.

3.2.2 Measures

3.2.2.1 Collective Intelligence Tasks

For each of the following tasks except RAPM, items were presented in a fixed order, and participants answered each question twice: first individually, then together with their teammate. This approach was employed so individual and collective decisions could be compared, enabling us to examine the two heads are better than one effect for each task. This repeated-measures approach is well-established in dyadic decision-making research (e.g., Bahrami et al., 2010; Blanchard et al., 2020; Koriat, 2015).

Applying Decision Rules (ADR; Bruine de Bruin et al., 2007). This test included 10 items that each presented 5 DVD players, their prices, and ratings (very low to very high) on 4 attributes (i.e., picture quality, sound quality, programming options, and reliability of brand). Participants were provided with a fictive customer's preferences on price and/or the 4 attributes and were required to select the DVD player that best matched their preferences. For example, participants were asked, "*LaToya only wants a DVD player that has got a 'Very*

High' rating on Reliability of Brand. Which one of the presented DVD players would LaToya prefer?" There was always one correct answer. Accuracy on this test is a mixed measure of Fluid Reasoning and Crystallised intelligence and possesses good internal consistency (.73). After each item, participants provided a confidence rating ranging from 20% (guessing) to 100% (completely certain).

Cognitive Reflection Test (CRT; Frederick, 2005; Toplak et al., 2014). This test included 7 items composed of numerical problems that elicited biased judgments because people tend to rely on a heuristic instead of conducting a simple mental calculation. For example, participants were asked, "*Together a bat and a ball cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?"* Responses were free text. Accuracy on this test is a measure of Quantitative Knowledge (Otero et al., 2022) which is a broad cognitive ability and has demonstrated good reliability (.72). After each item, participants provided a confidence rating ranging from 0% (guessing) to 100% (completely certain).

Geography test (Kleitman & Stankov, 2001). This test included 11 items that assessed participants' knowledge of Australian geography. Each item presented a question with two response options, and participants were required to select the correct one. For example, participants were asked, "*Which of the following states has a larger population: New South Wales or Victoria?*" Accuracy on this test is a measure of Crystallised Intelligence. After each item, participants provided a confidence rating ranging from 50% (guessing) to 100% (completely certain) and made a bet decision (Yes or No). The 11 items were selected from the 140 items used in the original study.

Ravens Advanced Progressive Matrices (RAPM; Raven, 1938-65). This test included 36 items. Each item displayed a 3x3 matrix of abstract figures that presented a horizontal and vertical pattern. The bottom right figure was blank, and participants were required to choose which of eight options completed the pattern. Accuracy on this test is a measure of Fluid

Reasoning. Internal consistency on this test has been shown to be excellent for accuracy (.80-81) and confidence (.90-.92; Blanchard et al., 2020; Blanchard et al., 2023). After each item, participants were asked to provide a confidence rating ranging from 12.5% (guessing) to 100% (completely certain) for the correctness of their response. The items from this test were split evenly to make 2 versions: individuals completed the 18 odd items, and dyads completed the 18 even items. This is known to produce equivalent short versions of RAPM.

3.2.2.2 Other Measures

Each of the following measures was completed by individual participants.

Composite Emotions Task (Wilhelm et al., 2014). This 36-item test assessed social sensitivity or one's ability to accurately perceive emotions. Each item presented a composite image composed of two photos of the same face expressing different emotions (e.g., the top half of the face displayed anger, and the bottom half of the face displayed happiness). Each composite image presents two of six emotions: sadness, disgust, fear, happiness, anger, or surprise. Participants were directed to identify the emotion shown on either the top or bottom half of the face and responded by selecting which of the six emotions were displayed in the target half of the face. Accuracy on this test has been shown to possess excellent reliability (.81).

Medical Decision Making Test (MDMT; Jackson & Kleitman, 2014; Jackson et al., 2016a, 2016b; Jackson et al., 2017). In this test, participants were told they were specialists in the Alpha virus, which can occur in multiple forms (regular or one of three mutations). They were given 3-minutes to memorise the pattern of associations between 9 symptoms and each form of the virus. For each of the 16-items, a fictive patient was presented with two symptoms. Participants diagnosed them as having the regular or one of the mutated forms of the virus. After each item, participants provided a confidence rating ranging from 25%

(guessing) to 100% (completely certain) and decided whether to treat the patient immediately or request a blood test to provide further diagnostic information. Patients survived if treated following a correct diagnosis but died if treated following an incorrect diagnosis. Blood tests resulted in a correct diagnosis and treatment, but only 50% of untreated patients survived while waiting for blood test results. The goal was to save as many lives as possible. Accuracy on this test is a measure of short-term memory and fluid reasoning and has demonstrated good reliability (.89-.92). In the current study, this test was used to compute Bias scores: positive scores indicated overconfidence, negative scores indicated under confidence, and scores approaching zero (±10 percentage units) indicated unbiased confidence ratings (Keren, 1991; Stankov et al., 2014; Yates, 1990).

Running Letter Span (Broadway & Engle, 2010; Kane et al., 2004; Pollack et al., 1959). For each trial, participants were instructed to recall the last *n* letters after seeing a sequence of individually appearing letters which flashed on their screen. They were not told how many letters would be shown in total and had to recall letters in the order they appeared. For example, they were instructed to remember the last 2 letters, and the sequence "*X Y T R S*" appeared. The correct answer was "*R S*". The number of letters to be recalled (*n)* ranged from three to seven, and sequences ranged from five to nine letters. The task contained five practice trials with feedback and 15 test trials without feedback. Accuracy on this test is a measure of Working Memory. Internal consistency estimates are excellent for accuracy on this test (.85).

Mini-IPIP (Donnellan et al., 2006). This questionnaire presented participants with 20 statements and asked them to rate the degree to which they were an accurate description of them using a five-point rating scale. For example, participants rated "*Am the life of the party*" from being a *very inaccurate* (1) to *very accurate* (5) description of them. This scale measures the Big-Five personality factors and has been shown to possess acceptable internal

consistency for Agreeableness (.70), Conscientiousness (.69), Extraversion (.77), Intellect (.65), and Neuroticism (.68).

3.2.3 Communication Measures

We recorded the conversations between group members while they completed the group tasks. Using these recordings, we computed the number of talking turns (frequency) to calculate equality of turn taking. Equality of turn taking, as defined by Woolley et al. (2010), refers to the standard deviation computed on the total number of talking turns for the members of a group. In the current study, zero indicated equality as both group members had the same number of talking turns, and higher values indicated greater inequality. To aid interpretation, we labelled this variable *inequality of turn taking.* This variable was computed separately for each test in the collective intelligence battery and overall.

3.2.4 Procedure

All participants were randomly assigned to dyads when they arrived at the university computer lab. Up to four participants (two dyads) completed the two-hour study at a testing session. Group members were seated at computers next to each other. Computer screens were arranged so that group members could see each other but not each other's screens. The order of tasks was counterbalanced to reduce the impact of practice or fatigue effects, and to prevent dyads completing the study at the same time from overhearing answers to the same task. After providing consent, all participants completed a demographic questionnaire, the cognitive tests, and Mini-IPIP. The same items were completed by individuals and dyads on the CRT and ADR tests. For these tests, participants answered a question alone then again with their teammate before moving onto the next question. When answering individually, the same item appeared on each group member's screen. Participants indicated a response using their keyboard. Participants then typed how confident they were that their answer was

correct. Participants were instructed to wait for their partner before proceeding to the group stage. When both group members were ready, they pressed the spacebar on their keyboards. The same item appeared on each member's screen accompanied by instructions to "*Discuss your answer with your partner. Try to persuade them if necessary. Come to an agreement and give the same answer."* After submitting the same answer, participants indicated how confident they were that their group answer was correct. They were instructed to answer this alone, without discussing their level of confidence with their partner. After submitting a confidence rating, the next question began. For RAPM and MDMT, participants completed matched versions of the tests as individuals and dyads. The two versions were completed as described above with an important difference: individual and group items were completed in separate blocks. The protocol was approved by the University of Sydney Human Research Ethics Committee (Project Number 2017/729).

3.3 Results

3.3.1 Descriptive statistics

3.3.1.1 Accuracy and Confidence

All analyses, except internal consistency estimates, were based on dyads as the unit of analysis. Thus, "individual" results refer to the average of the two group members working alone. This approach aligns with previous research on dyadic decision making (Bahrami et al., 2010; Bang et al., 2014; Blanchard et al., 2020; Koriat, 2015; Schuldt et al., 2017). The descriptive statistics and internal consistency estimates for the five measures of accuracy and confidence, for both individuals and dyads, are presented in Table 3.2. Omega total (McDonald, 1999) was used to measure internal consistency since we assumed unidimensionality but not tau-equivalence for each of the variables. The *t*-tests examined differences between individuals and dyads across each variable for each test.

Table 3.2. *Descriptive Statistics and Internal Consistency Estimates for Measures of Accuracy and Confidence (N=105)*

	Individuals			Dyads			
	ω_t	Mean	SD	ω_t	Mean	SD	*t*-value
Accuracy							
ADR	.71	58.33	18.78	.73	74.15	20.93	-16.00***
CRT	.75	47.65	23.41	.78	70.25	29.43	-16.43***
MDMT	.82	55.03	19.54	.80	86.43	16.84	-17.15***
RAPM	.83	68.88	15.68	.65	79.51	12.84	-10.96***
GT	.43	71.01	11.63	.42	75.00	13.48	-4.30***
Confidence							
ADR	.91	83.68	13.45	.92	88.61	11.66	-10.17***
CRT	.80	74.03	15.04	.81	83.69	13.83	-11.52***
MDMT	.97	63.83	18.38	.97	87.14	13.36	-15.52***
RAPM	.92	71.34	13.55	.92	76.77	13.32	-6.33***
GT	.93	72.72	9.76	.87	76.77	10.74	-9.34***

Note: ω_t = Internal consistency measured using Omega total; ADR = Applying Decision Rules; CRT = Cognitive Reflection test; MDMT = Medical Decision Making test; RAPM = Raven's Advanced Progressive Matrices; GT = Geography Test.
*** $p < .001$

The means and standard deviations for individual accuracy and confidence were comparable with other studies that used the same measures with undergraduate populations (Blanchard et al., 2020; Jackson et al., 2016a, 2016b; Jackson et al., 2017; Law et al., 2022). We also examined differences between individuals and dyads on accuracy and confidence. For each test, accuracy and confidence were higher for dyads than individuals. This is consistent with previous research, which found that groups tend to be more accurate (e.g., Bahrami et al., 2010; Henry, 1993; Hill, 1982; Tindale, 1989; Zarnoth & Sniezek, 1997) and confident (e.g., Blanchard et al., 2020; Koriat, 2015; Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997) than individuals across a range of cognitive and decision-making tasks.

For accuracy, internal consistency ranged from acceptable (.65) to excellent (.83) for individuals and dyads across all tests except the geography test which demonstrated poor internal consistency for individual (.43) and dyadic responses (.42). For confidence, internal consistency was excellent ranging from .80 to .97.

When we included the geography test in the CFA model, the loading (.21) and communality (.04) values were poor suggesting that accuracy on the geography test shared minimal variance with the underlying collective intelligence factor and did not fit well into the model. Thus, the geography test was removed from all subsequent analyses. Refer to Table B7, B8, and B9 in Appendix B for detailed results of this analysis.

3.3.1.2 Individual Difference and Communication Measures

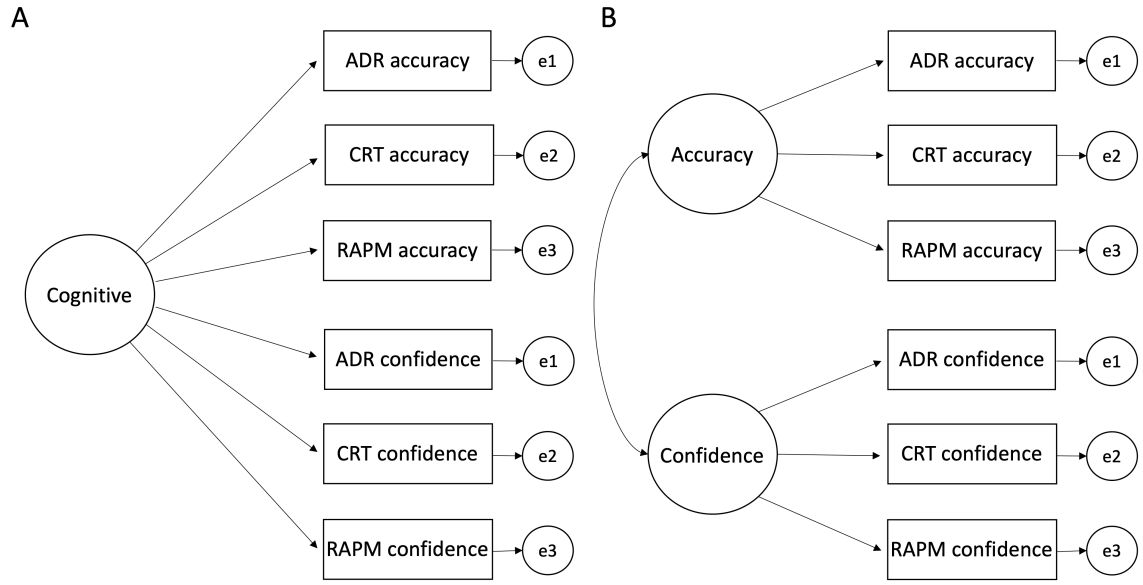
To maintain focus on the intelligence and confidence variables, descriptive statistics and internal consistency measures are presented in Tables B1 and B2 in Appendix B.

3.3.2 Extracting Collective Intelligence and Confidence factors

Using the same approach employed by Woolley et al. (2010), metrics of collective intelligence and collective confidence were estimated using CFA to extract latent factors from measures of accuracy and confidence recorded across the three collective intelligence tasks. We conducted CFA using the Maximum Likelihood method via the lavaan package (Rosseel, 2012) in R. We fitted and compared two first-order models that examined the factor structure of collective accuracy and confidence scores across the three tests for dyads. We tested hypothesized models with one first order factor and two first-order factors (see Figure 3.1). Each model was first tested without modification; thus, only hypothesized variables were allowed to define the respective factor, and correlations between latent constructs were freely estimated. Next, we tested modified versions of the same model to compare the fit indices. For these modified models, only the hypothesized variables were allowed to define the respective factor, but the error terms of accuracy and confidence derived from the same test were allowed to correlate because they were not independent and were derived from the same measure. See Figure 3.1 for diagrams of the unmodified one- and two-factor models that

were tested. Descriptions of the models, and a summary of fit indices, are presented in Table 3.3.

Figure 3.1. *Hypothesized One First-Order Factor Model (A) and Two First-Order Factors Model (B) Without Modification. Solid Lines Represent Positive Loadings/Correlations*



A modified two-factor model had the best fit for collective accuracy and confidence (model 3^{d}). In this model, the error terms of accuracy and confidence were correlated within the same test for RAPM and CRT. The Pearson correlation between accuracy and confidence was .46 ($p < .001$) for ADR, .66 ($p < .001$) for CRT, and .68 ($p < .001$) for RAPM. When the error terms were correlated across all 3 tests, the model was overfitted with Tucker-Lewis Index exceeding 1, so we decided to omit the variables from the test with the smallest correlation (i.e., ADR). The fit indices for this modified two-factor model were excellent: R^2 = .57; χ^2/df = 1.04; Goodness of Fit Index (GFI) = 0.99; Tucker-Lewis Index (TLI) = 0.99; Comparative Fit Index (CFI) = 0.99; Root Mean Square Error of Approximation (RMSEA) = 0.02 (collective intelligence = .00-.13). The results of this CFA model are displayed in Table 3.4. More detailed results of this model and the other two-factor models tested are presented in Tables B3, B4, B5, and B6 of Appendix B.

Table 3.3. *Summary of Fit Indices for Different Models of Collective Intelligence and Collective Confidence Using CFA (N = 105)*

Model	Fit Statistics									
	R^2	χ^2	*df*	X^2 / *df*	χ^2 diff	GFI	TLI	CFI	RMSEA (95% CI)	AIC
One-factor[a]	.51	69.46	9	7.72	-	0.99	0.68	0.81	.25 (.20-.31)	5014
Two-factor 1[b]	.57	57.88	8	7.24	11.58***	0.99	0.70	0.84	.24 (.19-.30)	5004
Two-factor 2[c]	.58	16.72	7	2.39	41.17***	0.99	0.93	0.97	.11 (.04-.19)	4966
Two-factor 3[d]	.57	6.23	6	1.04	51.65***	0.99	0.99	0.99	.02 (.00-.13)	4957

Note. GFI = Goodness-of-fit index; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; AIC = Akaike Information Criterion. The accepted model is in bold.

*** $p < .001$

[a]One-factor model consisted of one broad first order Cognitive factor defined by all the measures employed in the study without any modifications to the model.

[b]Two-factor model consisted of an intelligence factor (defined by all accuracy measures) and a confidence factor (defined by all confidence measures) without any modifications to the model.

[c]Two-factor model (intelligence and confidence factors) where error terms of the corresponding accuracy and confidence scores from RAPM were correlated.

[d]Two-factor model (intelligence and confidence factors) where error terms of the corresponding accuracy and confidence scores from RAPM and CRT were correlated.

The extracted factors were interpreted as follows:

Factor 1, *collective intelligence*: As hypothesized, this factor was defined by the loadings of all accuracy scores on the three tasks: ADR, CRT, and RAPM. These measures are all known to capture cognitive ability, thereby defining a collective intelligence trait.

Factor 2, *collective confidence*: As hypothesized, this factor was defined by loadings of all confidence scores on the three tasks. This factor captured a collective metacognitive confidence trait.

There was a strong, positive correlation between individual intelligence and collective intelligence ($r = .65$, $p < .001$), and individual confidence and collective confidence ($r = .70$, $p < .001$). There was also a strong correlation between the collective intelligence and collective confidence factors ($r = .85$, $p < .001$) which might suggest a one-factor model, however, the CFA model with one-factor had poor fit indices (see Table 3.3).

Table 3.4. *Summary of Standardised Regression Weights, Communalities, and Correlations for a CFA (N = 105)*

Measures	Intelligence	Confidence	h^2
ADR accuracy	.60	-	.35
CRT accuracy	.86	-	.73
RAPM accuracy	.65	-	.42
ADR confidence	-	.80	.63
CRT confidence	-	.93	.86
RAPM confidence	-	.68	.47
Factor intercorrelations			
Intelligence	-	.85***	

Note. All loadings and the factor intercorrelation were significant with $p < .001$

3.3.3 The Predictors of Collective Intelligence and Confidence

To examine our first three hypotheses, we conducted a hierarchical regression analysis. The independent variables included in the model were individual intelligence, individual confidence, social sensitivity, inequality of turn taking, the number of female members, working memory accuracy, and Big-5 personality[3]. The number of female members was treated as a categorical variable with three levels: all male dyads (baseline), mixed gender dyads, and all female dyads. Consistent with the approach employed by Woolley et al. (2010), individual metrics of intelligence and confidence were represented by mean accuracy and confidence scores measured with RAPM. Woolley et al.'s key variables were included in block 1, the control variables in block 2, and individual intelligence and confidence were added in the final block. To examine, hypothesis four, the same hierarchical regression model was fit with collective confidence as the outcome variable. See Table 3.5 for a summary of the results.

[3] To retain power in the regression analyses, missing values were imputed using multiple imputation via the mice package in R (van Buuren & Groothuis-Oudshoorn, 2011). One participant had non-random missing data for 6 of the 12 predictors, thus was removed. The remaining missing data were random with less than <1% for each predictor. For the focal variables, results were the same when analyses were conducted using the original and imputed data. See Table B10 in Appendix B for the results using data without imputation.

Table 3.5. *Hierarchical Regression Analyses Predicting Collective Intelligence and Collective Confidence (N = 105)*

	Collective Intelligence Block			Collective Confidence Block		
	1	2	3	1	2	3
Predictor	β	β	β	β	β	β
Mixed gender dyads	-.15	-.18	-.14	-.29	-.27	-.28
Female dyads	-.85**	-.82**	-.54*	-1.08***	-1.05***	-.64**
Social sensitivity	.39***	.34***	.22**	.28**	.20*	.13
Inequality of turn taking	.09	.08	.08	.14	.11	.12
WM accuracy	-	.23*	.10	-	.33***	.22**
Agreeableness	-	-.02	.03	-	.11	.13
Conscientiousness	-	-.01	-.04	-	.06	-.02
Extraversion	-	-.20*	-.09	-	-.01	.07
Intellect	-	.08	-.09	-	-.03	-.22**
Neuroticism	-	.01	-.04	-	.01	-.02
Intelligence	-	-	.33**	-	-	-.02
Confidence	-	-	.32***	-	-	.63***
R	.48	.59	.77	.50	.59	.79
R^2	.23	.35	.60	.25	.35	.63
ΔR^2	.23***	.12*	.25***	.25***	.10*	.28***

Note. WM = Working Memory. β = standardised regression coefficient.
*** $p < .001$, ** $p < .01$, * $p < .05$

Collective intelligence. In block 1, social sensitivity (β = 0.39, $p < .001$) and all female dyads (β = -0.85, $p < .01$) were significant predictors, accounting for 23% of the variance in collective intelligence. In block 2, working memory accuracy and the Big-5 personality traits together accounted for an additional 12% of the variance in collective intelligence ($\Delta R^2 = .12$, $p = .01$). In block 3, individual intelligence (β = 0.33, $p < .01$) and individual confidence (β = 0.32, $p < .001$) accounted for an additional 25% of variance in collective intelligence ($\Delta R^2 = .25$, $p < .001$). In support of hypotheses 1 and 3, individual intelligence and confidence were significant positive predictors of collective intelligence, and both were stronger predictors than social sensitivity, inequality of turn taking, and the gender

composition of dyads. Furthermore, in support of hypothesis 2a, female dyads had significantly lower collective intelligence than male dyads (β = -0.54, p = .01), however, there was no difference between male dyads and mixed gender dyads (β = -0.14, p = .47). Hypothesis 2b was partially supported: equality of turn-taking did not predict collective intelligence (β = 0.08, p = .24), but social sensitivity was a significant positive predictor (β = 0.22, p < .01).

Collective confidence. In block 1, social sensitivity (β = 0.28, p < .01) and female dyads (β = -1.08, p < .001) were significant predictors, accounting for 25% of the variance in collective intelligence. In block 2, working memory accuracy and the Big-5 personality traits together accounted for an additional 10% of variance in collective intelligence (ΔR^2 = .10, p = .04). In block 3, individual intelligence (β = -0.02, p = .87) and individual confidence (β = 0.63, p < .001) accounted for an additional 28% of the variance in collective intelligence (ΔR^2 = .28, p < .001). The significant predictors of collective confidence were individual confidence, the number of females, working memory accuracy, and intellect. In contrast, the significant predictors of collective intelligence were individual intelligence, individual confidence, social sensitivity, and the number of females. In support of hypothesis 4, the pattern of significant relationships differed for collective intelligence and collective confidence.

3.3.4 Mediation Analysis for the Proportion of Females, Social Sensitivity, and Collective Intelligence

A mediation analysis was conducted to examine whether social sensitivity mediated the relationship between proportion of females and collective intelligence. We conducted this analysis using the same data and predictors as block 3 of the hierarchical regression models.

We used nonparametric bootstrapping with 1,000 simulations to generate confidence intervals. The results are presented in Table 3.6.

Table 3.6. *The Results of a Mediation Analysis for Social Sensitivity Mediating the Relationship Between Proportion of Females and Collective Intelligence (N = 105)*

Effect		95% collective intelligence	
	Estimate	Lower	Upper
Indirect (ACME)	0.21**	0.05	0.42
Direct (ADE)	-0.54*	-0.89	-0.15
Total	-0.33	-0.69	0.06
Proportion Mediated	-0.62	-6.07	3.23

** $p < .01$; * $p < .05$

We found a significant positive indirect effect (ACME = 0.21, $p < .01$), indicating that social sensitivity significantly mediated the effect of the proportion of females on collective intelligence. This suggests that dyads with a higher proportion of females tended to have higher levels of social sensitivity which was associated with higher collective intelligence. We also observed a significant negative direct effect (ADE = -0.54, $p = .01$), suggesting that dyads with a higher proportion of females had lower collective intelligence, independent of social sensitivity. This was supported by small negative correlations between the proportion of females and dyadic accuracy on each of the tests used to measure collective intelligence: ADR ($r = -.11$, $p = .25$), CRT ($r = -.25$, $p < .01$), and RAPM ($r = -.10$, $p = .32$).

The total effect of the proportion of females on collective intelligence approached but did not reach significance (Total Effect = -0.33, $p = .11$), likely because the indirect and direct effects were in opposite directions and partially cancelled each other out. The proportion of the effect mediated was -.62 ($p = .11$) which also approached but did not reach significance. The wide confidence interval suggests instability in this estimate.

These results indicate that social sensitivity partially accounted for the relationship between the proportion of females and collective intelligence. However, the negative direct

effect between the proportion of females and collective intelligence suggests a complex relationship that depends on factors beyond social sensitivity.

3.3.5 Latent Profile Analysis

We then used individual and collective measures of intelligence, confidence, and bias to identify unique psychological profiles of dyads that clustered together on these measures. LPA was used to classify the observations under study into distinct profiles given their homogenous characteristics across a set of estimated values for the predictor variables.

3.3.5.1 Selecting an LPA Solution

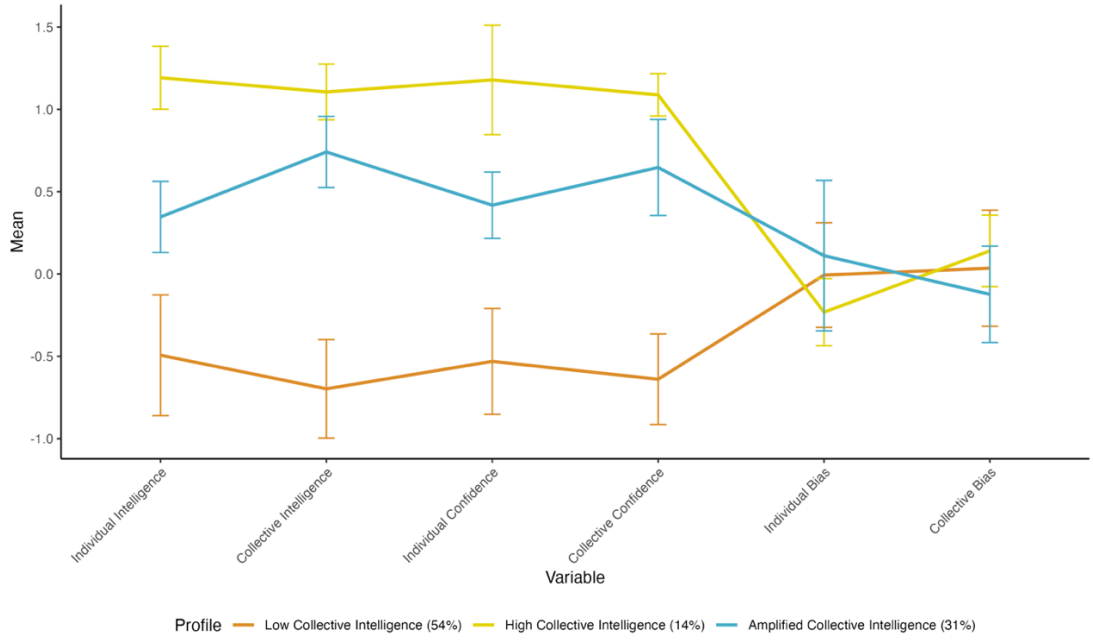
LPA was performed for solutions with 2-6 classes on six predictor variables. These variables were individual intelligence and confidence, the extracted factors for collective intelligence and collective confidence, and individual and collective bias scores. Goodness of fit statistics were used to identify the number of latent profiles (Clark & Muthén, 2009; Henson et al., 2007; Marsh et al., 2009; Spurk et al., 2020). Assessment of the indices and examination of the profiles within each model suggested a 3-Class solution was the best fitting model (See Figure 3.2). Refer to Table B11 and the corresponding text in Appendix B for a detailed summary of this assessment process.

3.3.5.2 Interpretation of the 3-Class Solution

The percentage of participants in each of the three classes was as follows: 54.29% in Class 1 (n = 57), 14.29% in Class 2 (n = 15), and 31.43% in Class 3 (n = 33). The three distinct profiles differed significantly on individual intelligence, collective intelligence, individual confidence, and collective confidence. The profiles were interpreted as follows: 1) *Low Collective Intelligence*: Those who were low on intelligence and confidence and well-calibrated at both the individual and collective levels; 2) *High Collective Intelligence*: Those who were high on intelligence and confidence and well-calibrated at the individual and

collective levels; and 3) *Amplified Collective Intelligence*: Those who were moderate on individual intelligence and significantly higher on collective intelligence, moderate on individual and collective confidence, and well-calibrated at both levels. Using the Bonferroni correction, Amplified Collective Intelligence had significantly higher collective intelligence than Individual intelligence ($t_{32} = 45.08$, $p < .001$). Furthermore, High Collective Intelligence had significantly higher collective Bias than individual Bias ($t_{14} = 3.35$, $p < .01$). There were no other differences between individual and collective scores within the profiles.

Figure 3.2. *Mean Scores for the Three Latent Profiles*



3.3.5.3 Differences Between the Three Profiles

First, a MANOVA was conducted to test whether the three profiles differed across social sensitivity, inequality of turn taking, the number of females, working memory accuracy, and big-five personality traits. A MANOVA indicated that the three profiles significantly differed across the theoretically relevant outcome variables ($F_{18,176} = 2.52$, Wilk's $\Lambda = .63$, $p < .01$, $\eta_p^2 = .21$).

Next, univariate ANOVAs were conducted to identify differences between the profiles on the relevant outcome variables. See Table 3.7 and Figure 3.3 for a summary of these analyses. The series of ANOVAs showed that the three profiles significantly differed on inequality of turn taking ($F_{2,96} = 3.11$, $p = .04$, $\eta_p^2 = .06$), the number of females ($F_{2,96} = 3.91$, $p = .02$, $\eta_p^2 = .08$), and working memory accuracy ($F_{2,96} = 8.43$, $p < .001$, $\eta_p^2 = .15$). Although, the differences, including the effect sizes, were moderate to small, with partial eta squared values ranging between .06 and .15.

Lastly, because there was an unequal number of members in each profile, Tukey-Kramer post-hoc tests were conducted on significant outcome variables to examine differences between the three profiles while controlling for multiple comparisons. Post-hoc tests revealed that the High Collective Intelligence profile had a significantly lower number of females ($p = .02$) and higher working memory accuracy ($p < .001$) compared with the Low Collective Intelligence profile. Furthermore, the Amplified Collective Intelligence profile had significantly higher inequality of turn taking ($p = .04$) and working memory accuracy ($p = .03$) compared with the Low Collective Intelligence profile. The High and Amplified Collective Intelligence profiles did not differ on any of the variables.

Figure 3.3. *Differences Between the Three Latent Profiles*

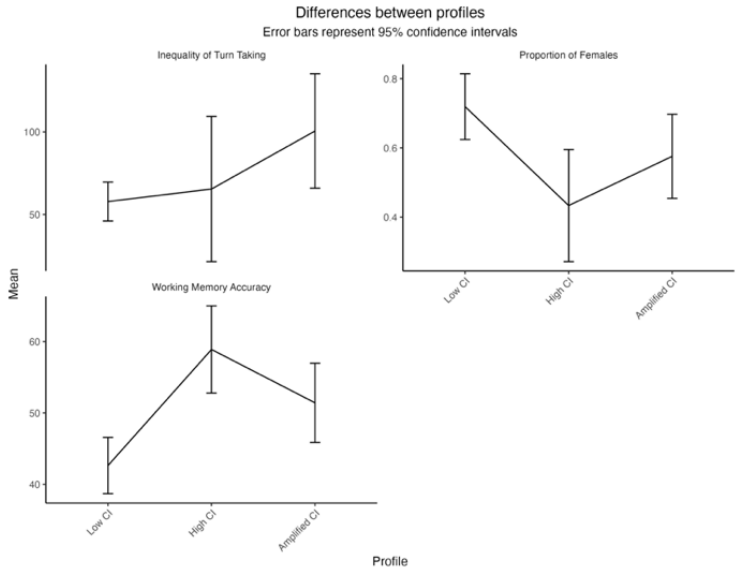


Table 3.7. *Results of ANOVAs for the Differences Between LPA Profiles (N = 105)*

Measure	Mean Low CI (SD)	Mean High CI (SD)	Mean Amplified CI (SD)	$F_{2,99}$	η_p^2	c1-2	c1-3	c2-3
Woolley et al. (2010) variables								
Social sensitivity	55.65 (10.26)	58.43 (8.73)	59.57 (9.20)	1.65	.03	.59	.20	.93
Inequality of turn taking	57.59 (45.77)	58.43 (8.73)	99.27 (103.06)	3.11*	**.**04	.58	**.04**	.32
Number of females	0.71 (0.37)	0.43 (0.32)	0.61 (0.34)	3.91*	**.08**	**.02**	.40	.28
Other variables								
Working memory accuracy	42.50 (15.27)	58.89 (12.06)	51.67 (16.04)	8.43***	**.15**	**<.001**	**.03**	.30
Agreeableness	4.02 (0.44)	3.91 (0.49)	3.92 (0.55)	0.58	.01	.70	.64	.99
Conscientious	3.24 (0.68)	3.13 (0.51)	3.17 (0.62)	0.23	.01	.82	.90	.97
Extraversion	3.03 (0.60)	2.69 (0.63)	2.92 (0.62)	1.92	.04	.14	.68	.49
Intellect/Openness	3.60 (0.52)	3.78 (0.60)	3.75 (0.60)	1.01	.02	.51	.49	.98
Neuroticism	2.96 (0.51)	3.18 (0.47)	2.91 (0.49)	1.59	.03	.26	.92	.21

Note. CI = collective intelligence. η_p^2 = partial eta squared. C = contrast. Contrasts were tested using the Tukey-Kramer post-hoc test. Significant contrasts and relevant eta squared values are in bold.

*** $p < .001$, ** $p < .01$, * $p < .05$

3.4 Discussion

The purpose of this study was to enhance the operationalisation of collective intelligence by employing a novel methodology for both measurement and analyses. Departing from previous research that relied on a theoretical taxonomy, the McGrath Task Circumplex, we used the empirically validated CHC model to guide the selection of well-structured tasks. This approach allowed us to assess the cognitive abilities relevant to collective intelligence. Our measure of collective intelligence captured three broad abilities: Fluid Reasoning, Crystallised Intelligence, and Quantitative Knowledge. We then examined the relationships between dyadic collective intelligence, individual intelligence, individual confidence, and other key characteristics. Additionally, we conducted a mediation analysis to investigate whether social sensitivity meditated the relationship between the proportion of females and collective intelligence. Finally, we utilised LPA to identify unique psychological profiles of dyads derived from measures of intelligence, confidence, and Bias. Finally, we assessed differences between these psychological profiles across several characteristics relevant to dyadic collective intelligence.

3.4.1 The Predictors of Collective Intelligence for Well-Structured Tasks

Our results demonstrate that individual intelligence is a strong predictor of collective intelligence in dyads for well-structured tasks. Furthermore, individual confidence also emerged as a strong predictor of collective intelligence, contributing uniquely beyond individual intelligence. These findings contrast with those of Woolley and colleagues using larger groups, who reported a weak correlation between individual intelligence and collective intelligence and emphasised the importance of social and compositional factors. By addressing previous methodological limitations, we have shown that smarter and more confident individuals form dyads with higher collective intelligence.

Currently, there is no consensus on how collective intelligence should be conceptualised or measured. Some researchers, such as Woolley et al. (2010), adopt a broad definition that includes a variety of task types and emphasizes emergent group processes such as coordination and communication. Graf-Drasch et al. (2022) proposed that different conceptualisations of collective intelligence may be required for well- and ill-structured tasks, guided by the findings that different cognitive and coordination processes are utilised for each task type. For well-structured tasks, general cognitive and coordination processes relate to performance, whereas ill-structured tasks may rely on task-specific processes (Newell & Simon, 1972; Schraw et al., 1995; Simon, 1973). Our approach, along with that of Rowe et al. (2024), builds upon this. We focus more narrowly than Woolley and colleagues on identifying the cognitive underpinnings of dyadic performance using well-structured tasks informed by models of individual intelligence. This approach allows for the systematic isolation and testing of how specific cognitive abilities contribute to group outcomes. It is not intended to provide a comprehensive account of collective intelligence but rather serve as a theoretically driven starting point for building a broader model of collective intelligence for well-structured tasks. Although our tasks required communication and coordination, thus capturing some key aspects of group interactions, we did not score these processes as outcomes, as in Woolley et al.'s work. We view these approaches as complementary rather than contradictory. Acknowledging these differing perspectives is essential for developing integrative models that specify how cognitive and interactional processes contribute to collective performance across different task contexts.

Our study focused on well-structured tasks, and we found that a single collective intelligence factor is appropriate for these tasks. In contrast, ill-structured tasks, which feature ambiguous pathways and multiple possible solutions, may require a multi-factor conceptualisation of collective intelligence as suggested by Graf-Drasch et al. (2022). We did

not assess collective intelligence for these tasks so we can only speculate that the presence of ill-structured tasks reduces a group's reliance on cognitive abilities and increases the importance of coordination processes for performance. This is supported by Woolley et al.'s (2010) results which are largely based on ill-structured tasks, where collective intelligence was predicted by factors related to coordination and social interaction (i.e., social sensitivity and equality of turn taking). Therefore, Woolley and colleagues' concept of collective intelligence may more accurately reflect teamwork processes and dynamics than the cognitive abilities driving group performance (Graf et al., 2019; Hackman & Morris 1975; LePine et al., 2008; Rowe et al., 2021). Ill-structured tasks are representative of many real-world scenarios encountered by groups; thus, understanding the task dependency of cognitive and coordination processes in collective intelligence is a critical question for future research.

There are several reasons why our findings diverge from those of Woolley et al. (2010). First, we addressed methodological limitations by using the empirically validated CHC model to guide task selection, rather than relying on theoretical taxonomies, such as the McGrath Group Task Circumplex, which may not adequately capture the cognitive abilities relevant to collective intelligence (Bell, 2007; Rowe et al., 2021). Second, by focusing exclusively on well-structured tasks, we avoided the confounding effects of mixing task types that require different cognitive and coordination processes. This allowed us to more precisely assess the role of individual intelligence in collective intelligence for well-structured tasks. Future research should examine collective intelligence in ill-structured tasks and a mix of well- and ill-structured tasks together. All are valid and important approaches, as many real-world projects are likely to contain subtasks that may blend both task types. Third, the difference in group size may have influenced the results. Our study, used dyads, whereas prior research typically involved groups with 2 to 5 members (e.g., Barlow & Dennis, 2016; Bates and Gupta, 2017; Engel et al., 2014; Rowe et al., 2024; Woolley et al., 2010). The

group dynamics that exist in dyads may emphasise cognitive abilities more than in larger groups, where coordination and interaction processes become more complex. Future research should investigate whether our findings apply to larger groups.

In addition to individual intelligence and confidence, we found that several other characteristics predicted collective intelligence. Consistent with Woolley et al. (2010), higher social sensitivity was associated with higher collective intelligence, and the proportion of females had a positive indirect relationship with collective intelligence through social sensitivity. We used a different measure of social sensitivity because the original test (Reading the Mind in the Eyes) has been shown to have several critical limitations (e.g., Black, 2017; Kittel et al., 2022; Olderbak et al., 2015). Thus, these findings support their claims that, the proportion of females and social sensitivity are important for collective intelligence beyond cognitive abilities. However, we also found a strong direct negative relationship between the proportion of females and collective intelligence. This indicates that, independent of social sensitivity, dyads composed of two females had lower collective intelligence scores than male only dyads but not mixed gender dyads. This divergence from Woolley et al. indicates that factors other than social sensitivity may be influencing the relationship between the proportion of females and collective intelligence in dyads performing well-structured tasks.

One possible explanation for this discrepancy lies in the cognitive demands of the tasks used to assess collective intelligence. Our tasks primarily assessed cognitive abilities, such as Fluid Reasoning and Quantitative Knowledge, that tend to confer a small performance advantage to males (Halpern et al., 2007; Irwing & Lynn, 2005; Otero et al., 2024). Supporting this, we found small to negligible negative correlations between the proportion of females and accuracy on each of the tests used to measure dyadic collective intelligence. These findings suggest that the cognitive abilities required by our well-

structured tasks may have contributed to lower collective intelligence scores in dyads with a higher proportion of females.

Another possibility is that the simpler social dynamics found in dyads compared to larger groups may reduce the influence of social sensitivity. In dyads, communication and coordination are more straightforward than in larger groups, and as a result these factors may play a diminished role in our study compared to Woolley et al. and others that used larger groups. This could also account for our failure to replicate the positive relationship between equality of turn-taking and collective intelligence. Furthermore, higher Working Memory accuracy was associated with higher collective intelligence, indicating that the capacity to hold and manipulate information in one's mind is important for dyadic collective intelligence on well-structured tasks.

As expected, the pattern of variables that predicted collective intelligence was different from those that predicted collective confidence. This finding supports Koriat's (2024) claim that measures of confidence provide unique information beyond accuracy. Confidence may reflect the consistency with which individuals or dyads would respond to the same task again in the future, not merely the correctness of their responses. This finding highlights the importance of considering confidence as a separate construct influencing dyadic performance.

3.4.2 The Psychological Profiles of Dyads

The application of LPA in our study allowed us to identify three distinct psychological profiles of dyads: 1) Low Collective Intelligence profile (approximately 54% of dyads); 2) High Collective Intelligence (approximately 14%); and 3) Amplified Collective Intelligence (approximately 32%). This person centred approach is novel in the study of collective intelligence and provides insights into how combinations of individual intelligence and

confidence contribute to dyadic performance. The High Collective Intelligence profile was comprised of dyads with two high individual intelligence and confidence members, leading to high collective intelligence. Furthermore, this profile had higher Working Memory accuracy and fewer females than the Low Collective Intelligence profile. The Amplified Collective Intelligence profile consisted of dyads where collective intelligence exceeded individual intelligence, suggesting that collaboration led to performance improvements. Interestingly, Amplified Collective Intelligence dyads had higher working memory accuracy and greater inequality of turn-taking compared to Low Collective Intelligence dyads, indicating that enhanced dyadic collaboration may involve strategic dominance by the more knowledgeable member. This contrasts with Woolley et al. (2010), who found that equality of turn taking predicted higher collective intelligence in larger groups. Our findings suggest that for dyads completing well-structured tasks, allowing the more competent member to lead may enhance collective performance. These results highlight the importance of considering individual intelligence, confidence, and possibly working memory and interaction patterns when forming dyads. Working memory accuracy and inequality of turn taking did not differ between High and Amplified profiles, so they may not offer comparable precision as intelligence and confidence.

3.4.3 Implications, Limitations, and Future Directions

Our findings have important implications for the measurement of collective intelligence, professional practice, and future research. Individual intelligence and confidence emerged as key predictors of collective intelligence on well-structured tasks. This suggests that forming dyads with smart and confident individuals may enhance dyadic performance without the need for a specific measure of collective intelligence. However, our findings are limited to dyads and may not generalise to larger groups.

As group size increases, Woolley et al.'s (2010) key predictors of collective intelligence (i.e., social sensitivity, equality of turn-taking, and proportion of females) may play a larger role as group dynamics like coordination and interaction processes become more complex. Future research should investigate collective intelligence and Woolley's key predictors using well-structured tasks and larger groups.

A specific measure of collective Intelligence, as developed by Woolley et al. (2010) may be more useful for ill-structured tasks which tend to rely on different cognitive and coordination processes than well-structured tasks. For these tasks, group interaction processes may be more important than cognitive abilities. Since many real-world tasks are ill-structured, future research should examine collective intelligence for these tasks separately. Preliminary research indicates that collective intelligence in ill-structured tasks may be represented by multiple factors that capture characteristics of coordination and interaction rather than cognitive abilities (e.g., Graf-Drasch et al., 2022; LePine et al., 2008). This raises an interesting question for future research: if coordination processes are more important for collective intelligence on ill-structured tasks, can they be trained to improve collective intelligence?

In professional settings, dyadic performance on well-structured tasks can be enhanced by selecting individuals with higher intelligence, confidence, and social sensitivity. Our findings suggest that organisations should consider incorporating cognitive and metacognitive assessments into their hiring and group formation processes to identify individuals who are likely to contribute effectively to dyadic collaborations.

Furthermore, the results of the LPA indicate that strategic dominance by the more knowledgeable or competent member may enhance performance in dyads with moderate individual intelligence. This suggests that dyads could be trained to identify and leverage the strengths of the more competent member. Training programs could focus on developing skills

for recognising expertise and allowing the more competent member to lead discussions and decision choices in a strategic way. It remains unclear whether members should assess competence at the task level (analogous to individual tests in our study) or at the project level (analogous to our battery of tasks). This presents an open question for future research which could improve dyad training and formation strategies.

Our findings should be interpreted alongside several additional limitations. While using three tests, as we did, is a recommended minimum to measure a latent trait like intelligence a larger number of tests could provide more robust measurement. We had a fourth measure, but the geography test demonstrated poor internal consistency and did not load on the collective intelligence factor when we fit a CFA model, thus it was removed from our analyses. The remaining tasks captured three of 16 broad cognitive abilities proposed under the CHC model of intelligence (e.g., McGrew, 2009). Future research should investigate collective intelligence using a larger battery of well-structured tasks that cover more of these broad abilities.

Our analyses were conducted at the dyad level, requiring us to transform individual-level variables, such as individual intelligence and social sensitivity, into a single score representing the two original scores of the dyad members. There are several methods for representing individual-level constructs at the dyad level, each with potential information loss that may be important for group dynamics: 1) the average score, which includes information from both members but reduces within-dyad variance; 2) the maximum score, which emphasizes one member's contribution but omits the other's; and 3) a difference score or similarity metric, which preserves within-dyad variance but discards the magnitude of the original scores. For our analyses, we used the average to represent all individual-level variables at the dyad level. While this approach, and the alternatives, produce trade-offs, our method is widely accepted in group decision-making research (e.g., Bahrami et al., 2010;

Bang et al., 2014; Blanchard et al., 2020; Koriat, 2015; Rowe et al., 2021; Schuldt et al., 2017). This issue was outside of the scope of this study. Future studies should systematically compare these and alternative methods to identify the best method of representing individual constructs at dyadic level.

An additional limitation is that we measured collective intelligence in naïve dyads who worked together for the first-time during our study. Prolonged collaboration may increase the impact of certain individual characteristics, such as personality traits, on dyadic dynamics. Moreover, the accumulated knowledge of individual members' abilities could moderate the relationships between key predictors and collective intelligence. Although, consistent with Woolley et al. (2010), we measured individual intelligence using a single test that captured Fluid Reasoning. This was adequate for our study, but future research should use a more comprehensive measure of individual intelligence. This may have the added benefit of clarifying the relationship between social sensitivity and collective intelligence.

3.4.4 Conclusion

Our study introduced a novel approach for examining dyadic collective intelligence on well-structured tasks by utilising the CHC model for task selection and incorporating metacognitive confidence. We found that individual intelligence and confidence are stronger predictors of dyadic collective intelligence than social sensitivity and interaction processes like equality of turn-taking. Notably, only one of Woolley et al.'s findings for the three key predictors (social sensitivity) was replicated in our study. These results challenge the generalisability of prior collective intelligence measures to dyads and highlight the importance of cognitive abilities and confidence in enhancing collective performance on well-structured tasks. Our findings have important implications for the selection and training of dyads, suggesting that focusing on individual intelligence, confidence, and social sensitivity may be effective at improving dyadic performance.

Chapter 4: Study 3
How Trait Confidence and Communication Shape Dyadic Decision Outcomes and Confidence Matching

The original manuscript for the study described in this chapter has been submitted to the journal *Cognitive Research: Principles and Implications* and is currently under review.

4.1 Introduction

4.1.1 Collective Decisions

How individuals share and integrate information when working in a group is critical for success (Hastie & Kameda, 2005). Collective decisions are often characterised by uncertainty and choices between known options. For example, a driver and navigator must assess terrain features to determine the most efficient route, or a pair of intelligence analysts may evaluate conflicting reports to reach a consensus on the likelihood of enemy activity in a specified area. In such contexts, achieving a "two heads are better than one" effect depends not only on pooling information, but also on how effectively they communicate, align confidence in their judgments, and integrate perspectives to arrive at a joint decision. For dyads, simply following the majority is not viable, as disagreement leads to a tie. This raises an important question: how do dyads resolve disagreement and achieve a collective benefit?

Bahrami et al. (2010) provided an answer to this question and an important extension on classic small group research (e.g., Festinger, 1954; Hill, 1982; Hinsz, 1990; Sniezek & Henry, 1989; Tindale, 1989) by showing that dyads achieve a collective benefit by sharing and using each other's confidence. This strategy works because higher-confidence ratings often signal a higher probability of being correct (Koriat, 2008, 2012b; Stankov & Crawford, 1998; Yaniv, 1997). Over repeated trials, dyad members' confidence ratings tend to shift towards each other, an alignment process that has been called "confidence matching" (Bang

et al., 2017). Confidence matching is necessary because two individuals can differ in their baseline levels of trait confidence, and this can undermine their ability to make correct decisions. For example, if one member always rates their confidence between 50-70% and another between 80-100% then the dyad's final answers are likely to reflect the latter's higher-confidence opinions. This scenario would be beneficial if the higher-confidence member had superior ability, but that isn't always the case (Blanchard et al., 2020). Several studies have further examined the role of confidence matching in collective decision-making (e.g., Friedemann et al., 2024; Schneider et al., 2024), including Pescetelli and Yeung (2022) who extended Bang et al. (2017) finding by demonstrating that confidence matching is associated with improvements in decision accuracy under more natural conditions.

Working in pairs or small groups also boosts overall confidence for a task (Patalano & LeClair, 2011; Savadori et al., 2001; Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997) and this rise can occur even when decision accuracy does not improve (Blanchard et al. 2020; Heath & Gonzalez, 1995; Minson & Mueller, 2012; Schuldt et al., 2017). Two recent studies reported that the size of a dyad's increase in task-relevant confidence (decision confidence) depends on the trait confidence levels of its members, suggesting that trait confidence plays an important role in how decision confidence develops within dyads.

Decision confidence refers to confidence judgments made for responses to items within a specific task. These ratings reflect moment-to-moment monitoring of one's performance and guide collective decisions in real-time (e.g., Koriat, 2008). Trait confidence, by contrast, represents a stable, domain-general tendency for confidence judgments across different tasks and contexts, relative to others (e.g., Johnson, 1939; Kleitman & Stankov, 2001). The two constructs are related as trait confidence is typically derived from multiple decision confidence measures taken across different cognitive domains (Stankov et al., 2014). As such, decision confidence captures situational judgments, while trait confidence reflects

broader individual differences in metacognitive self-beliefs. Crucially, trait confidence is embedded within decision confidence ratings such that every decision confidence judgment carries both task-specific and person-specific information. This overlap suggests that trait confidence may shape how decision confidence develops within dyads and influence how individuals respond to their partner's judgments during collaboration.

The present study focused on trait confidence as a potential moderator of how dyads share information and improve their decisions. Specifically, we investigated whether baseline levels of trait confidence can influence the extent to which collective decisions and decision confidence improve compared to individual decisions. Furthermore, we examined whether trait confidence moderated the development of confidence matching and its relationship with dyadic accuracy gains. By examining the dyadic processes and outcomes of individuals with high-trait versus low-trait confidence, we aimed to provide a more nuanced account of when and why "two heads are better than one."

4.1.2 The Confidence Theory

Bahrami's et al.'s (2010) influential study of the "two heads are better than one" effect highlighted the critical role of subjective confidence in dyadic decision-making. In their study, participants viewed two visual stimuli, each containing six identical patterns. Within one stimulus, a single pattern had increased luminance, creating a subtle target for participants to detect. Participants first judged which stimulus contained the target and provided a confidence rating individually and then again with a partner, during which dyad members freely discussed and agreed upon a joint decision. Results demonstrated that joint responses were more accurate than individual responses. Among multiple competing explanations, the most compelling was that dyad members shared and used each other's subjective confidence as a *cue* for decision accuracy – a process we refer to as the confidence theory. Given that decision confidence typically correlates positively with decision accuracy

(Koriat, 2008; Stankov et al., 2014; Yaniv, 1997), relying on confidence judgments is an effective strategy. Subsequent studies by Koriat (2012a, 2015), provided robust support for the confidence theory, demonstrating that dyad members continue to rely on confidence judgments even when they provide a misleading signal, producing a "two heads are worse than one" effect. The confidence theory assumes that individuals can reliably interpret each other's confidence ratings, such that a response associated with greater confidence is more likely to be correct. An important extension of this theory is whether dyad members adapt their own confidence to better align their ratings with their partner's. This process is known as confidence matching.

4.1.3 Confidence Matching

Bang et al. (2017) demonstrated that confidence matching is a heuristic strategy that dyads use to negotiate influence on joint decisions. They employed the same perceptual discrimination task as Bahrami et al. (2010) and participants completed the study under two conditions. In the social condition, each participant first answered individually. Then, their initial response and confidence rating were displayed on their partner's computer screen, and the response associated with higher confidence was automatically selected as the joint decision. In the isolated condition, participants completed the task individually without any interaction. Across six experiments and at least 160 trials per condition, participants in the social condition shifted their confidence ratings towards each other across trials. That is, confidence ratings converged towards a shared scale. A phenomenon termed confidence matching. Confidence matching occurred both with and without explicit feedback, although the effect was stronger when feedback was provided. These findings indicate that confidence matching functions as a strategy to influence joint decisions. Importantly, Bang et al. (2017) found that this strategy maximized decision accuracy when dyad members had similar ability but produced smaller accuracy gains when their ability differed. The latter effect occurred

because confidence matching caused the more competent member's judgments to be weighted less and the less competent member's judgments to weighted more. That is, the less competent member's judgments were endorsed more often than justified by their probability of being correct. Nonetheless, even under differing conditions, dyads generally remained more accurate than individuals.

Bang et al.'s (2017) approach explicitly incentivised participants to engage in this confidence matching behavior through the use of an automated decision rule. This raises the question of whether confidence matching naturally emerges in situations more analogous to real-world settings where dyads must interact before making decisions. To explore this issue, Pescetelli and Yeung (2022) used a perceptual task with 432 trials split across three phases. Phases one and three were completed individually, while the second phase allowed non-verbal social interaction between dyad members. During the social phase, participants saw their partner's initial response and confidence rating on their computer screen and could observe their partner updating their responses in real-time before submission. Unlike Bang et al. (2017) participants did not make joint decisions, nor did they verbally communicate. Pescetelli and Yueng (2022) observed rapid confidence matching during the social phase that was present almost immediately. Confidence matching was also associated with a small improvement in decision accuracy following interaction.

Extending Bang et al.'s (2017) theory of confidence matching, Pescetelli and Yeung (2022) suggested that confidence matching facilitates more accurate sharing of task-specific variance in confidence (decision confidence) while simultaneously reducing the influence of task-irrelevant variance (which they labelled as trait-confidence). They argued that trait confidence is influenced by domain-general factors such as socio-economic background, profession, and personality and that these influences obscure performance-based confidence signals. Although, Pescetelli and Yeung (2022) viewed trait confidence largely as noise or

irrelevant variance, it is important to consider whether trait confidence truly lacks informative value. Trait confidence may reflect meaningful individual differences that dyad members implicitly rely upon when making joint decisions.

4.1.4 The Role of Trait Confidence

Regardless of the specific cognitive tests used to capture confidence ratings (e.g., general knowledge, syllogistic reasoning, perceptual discrimination), a broad trait confidence factor emerges that is distinct from cognitive ability, which is defined by accuracy scores. That is, relative to others, some individuals consistently report low confidence across tests, while others consistently report high confidence. This robust, trait-like tendency has been replicated in numerous studies (e.g., Johnson, 1939; Kleitman & Stankov, 2001, 2007; Pallier et al., 2002; Soll, 1996; Stankov, 1999; Stankov et al., 2012a; Stankov et al., 2012b). Moreover, considerable evidence indicates that individuals generally possess good calibration across diverse cognitive tests, accurately aligning their confidence with their actual performance levels (Lichtenstein & Fischhoff, 1977; Koriat, 2012c; Nietfeld et al., 2005).

While trait confidence primarily reflects domain-general self-monitoring of knowledge and performance, it does have a small positive relationship with the personality trait Openness to Experience (see Stankov, 1999 for a review) and females tend to have better calibrated confidence than males (see Stankov et al., 2014 for a review). However, demographic and personality typically have minimal effects on shaping trait confidence relative to broader metacognitive processes.

Recent evidence suggests that decision confidence reflects the likelihood of response replicability rather than response accuracy (Koriat, 2024). Specifically, high-confidence judgments indicate that an individual is likely to repeat their initial response when presented with the same question again, even if that response is incorrect. Consequently, high-trait

confidence individuals may demonstrate greater stability and decisiveness at the cost of lower flexibility in revising erroneous judgments. This finding has implications for dyadic decision-making contexts, especially where dyad members first form a judgment individually and are then required to form a joint judgment for the same item or stimulus. To illustrate this practically, consider a real-world scenario where two intelligence analysts must collaboratively assess a set of ambiguous and time-sensitive intelligence reports to determine the likelihood of enemy movement in a given area. If both analysts possess high-trait confidence, they may quickly agree on a decisive interpretation and course of action, showing strong commitment to their initial judgments, even when faced with new or contradictory information. Conversely, if both have low-trait confidence, they may extensively question and reconsider each other's assessments, potentially leading to indecision or delayed responses. In a mixed-trait confidence dyad, the higher-trait confidence analyst may dominate the decision-making process, strongly advocating for their interpretation while potentially overlooking critical insights from their lower-confidence partner, who remains more open to alternative viewpoints. Thus, trait confidence levels may directly shape the interpersonal dynamics of collaborative decision making, influencing both the process and the quality of the outcome. The dynamic in mixed-trait confidence dyads may limit the potential decision-making advantages that often arise from collaboration, particularly among similarly capable individuals.

This inflexibility among high-trait confidence members may partly explain the observed asymmetry in confidence matching within dyads, where the lower-confidence member typically increases their decision confidence more than the higher-confidence member decreases theirs (Schneider et al., 2024). This asymmetry may inflate overall dyadic confidence, potentially increasing miscalibration between dyadic decision confidence and accuracy. This could account for the widely observed phenomenon that dyads typically report

higher-decision confidence than individuals (Patalano & LeClair, 2011; Savadori et al., 2001; Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997), often independent of a corresponding improvement in decision accuracy (Blanchard et al., 2020; Heath & Gonzalez, 1995; Minson & Mueller, 2012; Schuldt et al., 2017). Specifically, Mahmoodi et al. (2013) found that dyads increased more in confidence when they communicated verbally compared to non-verbally.

Two recent studies have directly examined how decision confidence changes as a function of the trait confidence composition of dyads. Controlling for cognitive ability, Schuldt et al. (2017) paired individuals with high-trait and low-trait confidence to form dyads of either high-trait, low-trait, or mixed-trait confidence. They found that dyads consisting of two low-trait confidence members had the largest increases in decision confidence. In contrast, Blanchard et al. (2020) found the opposite: dyads with higher-trait confidence showed the largest increases in decision confidence. Both studies used items from the same general knowledge pool. They also reported that increases in decision confidence did not correspond with improved decision accuracy; dyadic and individual decision accuracy was comparable across all conditions. The validity of the confidence theory depends on decision confidence reliably tracking decision accuracy. The conflicting findings might stem from differences in measuring trait confidence. Schuldt et al. relied on confidence ratings from a single alternate version of their dyadic task which notably lacked a correlation between decision confidence and accuracy. This limits its validity as a stable, domain-general measure of trait confidence which is supposed to monitor performance. In contrast, Blanchard et al. employed multiple, well-validated cognitive tests consistent with established trait confidence research (e.g., Kleitman & Stankov, 2007; Stankov et al., 2015; Stankov & Crawford, 1997), providing a more robust measurement of stable individual differences in trait confidence.

Despite its evident relevance to dyadic processes and outcomes, the influence of trait confidence on dyadic decision-making remains understudied. Our research directly addressed

this gap by investigating how trait confidence moderated changes in decision accuracy, changes in decision confidence, and confidence matching for dyadic decision making.

4.1.5 Profiling Dyadic Decision-Making

Most research on dyadic decision-making has taken a variable-centred approach, treating relationships between individual traits and group performance as consistent across all dyads. While informative, this approach overlooks meaningful heterogeneity in how dyads integrate individual differences to shape collective outcomes. For instance, studies such as Woolley et al. (2010) assumed homogeneous effects of predictors across dyads, potentially masking distinct subtypes of dyadic functioning.

To address this gap, we adopted a person-centred approach using LPA to identify distinct psychological profiles of dyads based on their levels of cognitive ability and trait confidence which are known to influence dyadic performance. This approach enabled us to identify subgroups within the sample who show distinct configurations of individual and collective-level traits and outcomes, offering a more nuanced view of how dyadic decision-making develops.

In study 2 (Chapter 3), LPA revealed three distinct psychological profiles of dyads based on individual cognitive ability and trait confidence, and their collective performance. The Low Collective Intelligence profile (54%) included dyads with low cognitive ability and trait confidence who underperformed collectively. The High Collective Intelligence profile (14%) consisted of dyads with high cognitive ability and trait confidence who achieved strong collective performance. The third group, Amplified Collective Intelligence (32%), included dyads whose collective performance exceeded the individual abilities of their members, suggesting a genuine collaborative benefit. This group was characterized by moderate individual scores, and superior collective performance. Compared to the other

profiles, they had greater working memory accuracy and more unequal turn-taking which facilitated strategic dominance when making collective decisions. These patterns suggest that, for well-structured tasks, asymmetric participation, where the more capable member takes the lead, may enhance dyadic outcomes, challenging prior assumptions that equal participation is universally beneficial.

These findings highlight the utility of LPA in revealing distinct patterns of collaboration that variable-centred methods may obscure. Specifically, they show that cognitive ability and trait confidence are key ingredients in determining whether dyads simply combine, amplify, or dilute their members' individual potential. Building on this framework, the current study applied LPA to examine whether similar latent profiles emerged in a different collaborative decision-making context, and whether these profiles related to communication and performance outcomes.

4.1.6 The Present Study

The present study aimed to replicate and extend previous findings on confidence matching and changes in dyadic confidence. We conducted a pre-screening study to measure trait confidence and cognitive ability, subsequently inviting individuals with high-trait or low-trait confidence to participate in the main study. Participants were paired into high-trait, low-trait, or mixed-trait confidence dyads.

To ensure methodological consistency with previous research on the confidence theory (Bahrami et al., 2010) and confidence matching (Bang et al., 2017; Pescetelli & Yeung, 2022) and to explore its interaction with trait confidence, we included a communication manipulation by varying the type of communication within dyads. In the isolated condition, participants responded to all items twice individually, without any interaction. In the passive condition (similar to Pescetelli & Yeung, 2022), participants

initially responded individually and then saw their partner's answers on their computer screen before making a second individual response. In the active condition, participants first responded individually, then verbally discussed their answers with their partner before jointly deciding on a response.

Verbal interactions allow dyad members to actively question their partner's reasoning and clarify uncertainties, thereby facilitating a deeper understanding of the underlying rationale behind a response. This richer exchange may enable dyads to better discriminate between responses that are held with higher confidence and are erroneous and those that are held with lower confidence but are accurate. By comparison, the passive viewing condition limited interaction, forcing members to rely exclusively on interpreting numeric confidence ratings without clarification, potentially restricting the precision and accuracy of judgments.

First, we examined whether trait confidence moderates the size of decision accuracy improvements when dyad members transition from individual to dyadic responses, controlling for cognitive ability. Prior studies using a similar design (Blanchard et al., 2020; Schuldt et al., 2017) did not find decision accuracy gains in dyads, possibly because confidence did not reliably signal accuracy in their general knowledge tests. In our study, assuming a positive correlation between decision confidence and accuracy, we predicted complex interactions between trait confidence, type of communication, and decision accuracy gains.

For mixed-trait confidence dyads, which were characterised by substantial differences in members' trait confidence, we expected decision accuracy gains in both the passive and active communication conditions. According to confidence theory, effective collaboration requires accurate assessment and use of each other's subjective confidence (Bahrami et al., 2010; Bang et al., 2017). However, without explicit feedback like in our study, large initial differences in trait confidence might limit decision accuracy gains more in the passive

condition where the transmission of social information is limited and the high-trait confidence member is likely to be more influential (Kerr & Tindale, 2004). The active condition, with verbal discussion and a consensus requirement, should facilitate greater decision accuracy improvements.

H1a: Mixed-trait confidence dyads will show decision accuracy improvements in both passive and active communication conditions.

H1b: Mixed-trait confidence dyads will have greater decision accuracy improvements in the active compared to the passive *communication* condition.

For dyads with matched levels of trait confidence (both low or both high), we expected accuracy improvements in both passive and active communication conditions because their initially aligned confidence scales should facilitate effective discrimination between response options. Differences were expected based on the consistency of responses and the flexibility of error correction (Koriat, 2024). High-trait confidence individuals may show strong decisiveness and resistance to revising their initial judgments, potentially suppressing the effectiveness of collaboration in the passive condition where they were not required to make joint decisions. Conversely, low-trait confidence dyads may approach interactions with greater openness to new information and more frequent reassessment of initial responses. Thus, low-trait confidence dyads should consistently benefit across both communication conditions.

H1c: Low-trait confidence dyads will improve decision accuracy equally in passive and active communication conditions.

High-trait confidence dyads, characterised by greater decisiveness but reduced flexibility in changing responses when presented with social information, should show limited accuracy gains in the passive condition, where the revision of responses is voluntary. By contrast, the active condition's requirement for consensus would encourage interrogation

of initial judgments, increased flexibility, and the correction of errors. This should produce larger decision accuracy gains.

H1d: High-trait confidence dyads will show decision accuracy improvements in both passive and active communication conditions.

H1e: High-trait confidence dyads will have greater decision accuracy improvements in the active compared to the passive communication condition.

Second, we expected to replicate Blanchard et al.'s (2020) finding that trait confidence moderates the size of a dyad's increase in decision confidence. Following Mahmoodi et al. (2013), we expected greater decision confidence increases in the active communication condition, independent of decision accuracy gains as the mere presence of social information appears to boost decision confidence (e.g., Heath & Gonzalez, 1995).

H2a: High-trait confidence dyads will show greater increases in decision confidence compared to low-trait and mixed-trait confidence dyads in both passive and active communication conditions.

H2b: Increases in decision confidence will be greater in the active compared to the passive communication condition.

Third, we tested whether confidence matching would emerge during verbal communication (i.e., active condition). Bang et al. (2017) and Pescetelli and Yeung (2022) demonstrated confidence matching using explicit numeric confidence ratings under communication conditions similar to our passive condition. Fusaroli et al. (2012) showed that there was a large amount of heterogeneity in expressions of confidence for verbally communicating dyads. Dyad members rarely communicated their confidence using numerical values, instead using phrases such as "I'm absolutely sure" or "that was a wild guess." Their findings showed that expressions of confidence became more similar over time within dyads and greater linguistic alignment of expressions of confidence was

associated with greater gains in decision accuracy. These findings provide indirect evidence that confidence matching may occur naturally in verbally communicating dyads.

We expected confidence matching to be more pronounced in the passive condition. Unlike verbal interactions, which rely on subjective interpretation of diverse verbal expressions of confidence, the passive condition presented confidence as explicit numeric values, making it easier for dyad members to quickly align their confidence judgments and reduce ambiguity about each other's subjective confidence levels.

H3a: Confidence matching will occur in both passive and active communication conditions.

H3b: Confidence matching will be stronger in the passive compared to the active communication condition.

Fourth, we examined how confidence matching relates to decision accuracy improvements. Based on Pescetelli and Yeung (2022) and Fusaroli et al. (2012), we expected a positive relationship between confidence matching and a dyad's accuracy gain for both passive and active communication conditions, but we expected a stronger relationship in the passive condition where social information was limited to numerical confidence ratings.

H4a: Confidence matching will positively predict a dyad's decision accuracy improvement in both passive and active communication conditions.

H4b: The relationship between confidence matching and a dyad's decision accuracy gain will be stronger in the passive compared to the active communication condition.

Finally, we adopted a person-centred approach to identify distinct psychological profiles of dyads based on individual cognitive ability and trait confidence. Building on our previous work using LPA (Chapter 3), which identified three distinct dyadic profiles (Low,

High, and Amplified Collective Intelligence), we aimed to replicate these findings and extend them by identifying psychological profiles across communication conditions.

Our prior findings revealed that the collective performance of a small number of groups exceeded individual capabilities. Dyads in the Amplified Collective Intelligence profile performed better as a group than would be predicted by the ability of their individual members. This amplification was associated with higher working memory and greater inequality in turn-taking, suggesting that asymmetrical collaboration, where the more competent member strategically dominates, may support collective gains.

In this study, we used LPA to assess whether similar psychological profiles would emerge under varying communication conditions. Our aim was to determine whether different combinations of individual cognitive ability and trait confidence would cluster into distinct profiles under different communication conditions. We also aimed to explore whether an amplified profile would emerge, similar to our previous study, where dyadic performance exceeded their individual abilities.

H5a: LPA will identify several latent profiles of dyads characterised by distinct patterns of individual cognitive ability and trait confidence, and distinct patterns of collective performance.

H5b: At least one profile will demonstrate amplified performance, where dyadic performance exceeds the individual abilities of its members.

4.2 Method

To examine the role of trait-confidence and type of communication on dyadic outcomes, we first developed 3 matched versions of a general knowledge test then we ran a pre-screening study on a large sample of individuals ($N = 1189$) to identify those with high-trait or low-trait confidence to invite for inclusion in the main study ($N = 210$). Our selection

criteria aimed to recruit individuals who scored beyond ±0.50 standard deviations on trait confidence and within ±1.50 standard deviations of the mean on cognitive ability. To keep the focus on the final stage of this research, descriptions of the general-knowledge test development and the pre-screening study are located in Appendix C. All stages of the study were approved by the University of Sydney Human Research Ethics Committee (Project Number 2019/706).

4.2.1 Main Study

The main study paired participants together based on their level of trait confidence measured in the pre-screening study and had them complete 3 matched versions of a general-knowledge test each under different communication conditions.

4.2.1.1 Participants

Participants were 210 Australian psychology undergraduates (52 males; Mean age = 22.02, SD = 6.12) who completed the study as 105 dyads. All participants previously completed the pre-screening study and half were identified high-trait confidence and the other half as low-trait confidence. They received either partial course credit or financial reimbursement for completing the study.

4.2.1.2 Measures

General Knowledge tests. The items for the general-knowledge test were sourced from several previous studies (Brewer & Sampaio, 2012; Blanchard et al. 2020; Schuldt et al., 2017; Stankov, 1997). Each version was originally composed of 22 two-alternative forced-choice items; however, due to poor reliability for decision accuracy, they were reduced post-hoc to 10 items. The remaining items covered a broad range of content areas: geography, art, music, film, history, science, and vocabulary. For example, *What does the word orthodox mean? Religious or Conventional** (* indicates the correct answer). The resulting internal

consistency estimates for each version were around .60. After each item, participants were asked to provide a confidence rating ranging from 50% (guessing) to 100% (completely certain) for the correctness of their response. The internal consistency estimates for confidence were around .80. The three versions were matched on decision accuracy, decision confidence, and content domain. Refer to Appendix C for more information about the general knowledge tests (see Tables C1, C2, C3, and C4).

Esoteric Analogies Test (EAT; Stankov, 1997). This test was administered in the pre-screening study. It involved participants completing 20 analogies. For each analogy, participants were presented with a pair of words and asked to select one of four options that reflected the same relationship with a target word. For example, *LOVE is to HATE as FRIEND is to: (1) LOVER, (2) PAL, (3) OBEY, (4) ENEMY**. Accuracy requires both reasoning skills and prior knowledge thus it is a mixed measure of Fluid Reasoning and Crystallised Intelligence. Prior research with Australian undergraduate samples reported acceptable internal consistency for decision accuracy (ranging from .69 to .74) and excellent internal consistency for confidence (ranging from .88 to .94; Jackson et al., 2016a, 2016b; Law et al., 2022).

Raven's Advanced Progressive Matrices (Raven, 1938-65). This test was administered in the pre-screening study. It consists of 36 items, each featuring a 3x3 grid of abstract figures forming a horizontal and vertical pattern, with the bottom right figure missing. Participants select one of eight possible options to complete the matrix. Accuracy is a measure of Fluid Reasoning. The internal consistency is excellent for decision accuracy, ranging from .80 to .81, and confidence, ranging from .90 to .92 (Blanchard et al., 2020; Blanchard et al., 2023). After responding to each item, participants rated their confidence in their answer on a scale from 12.5% (guessing) to 100% (completely certain). In the present study, participants completed a short 15-item version.

Mini-IPIP (Donnellan et al., 2006). This questionnaire was administered in the pre-screening study. Participants were presented with 20 statements about their personality, which they rated on a five-point scale ranging from “very inaccurate” (1) to “very accurate” (5). For example, one item asked participants to rate the accuracy of the statement “Am the life of the party.” This scale assesses the Big Five personality traits and has been found to have acceptable internal consistency for Agreeableness (.70), Conscientiousness (.69), Extraversion (.77), Intellect (.65), and Neuroticism (.68).

Several additional individual differences measures were administered to participants as potential covariates, including behavioural inhibition, empathy, motivational traits, psychological safety, risk aversion, social sensitivity, and trust within dyads. The trait confidence conditions did not differ significantly on any of these measures; thus, none were included as covariates in the main analyses. Detailed descriptions of these individual difference measures and the between-group comparisons are provided in Appendix C (see Table C5).

4.2.1.3 Communication Measures

Conversations between group members were recorded during the active communication condition. From these recordings, we calculated the number of words spoken, number of speaking turns, the equality of words spoken, and the equality of turn-taking. Following Woolley et al. (2010), equality of turn-taking (and equality of words spoken) was measured by computing the standard deviation for the total number of speaking turns for both dyad members. A value of zero indicated perfect equality, where both members contributed an equal number of turns (or words), while higher values reflected increasing levels of inequality. For clarity, we referred to this measure as *inequality of turn-taking*.

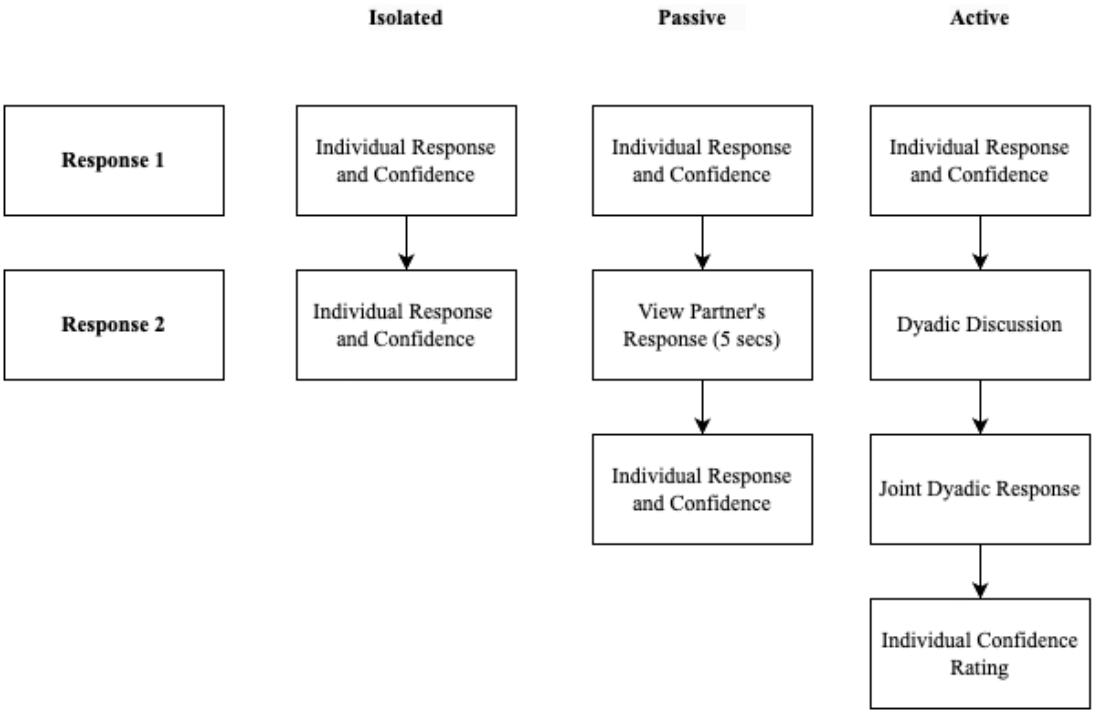
4.2.1.4 Procedure

Each session involved two participants forming a dyad, who completed the two-hour study online via Zoom. Participants were required to keep their cameras and microphones on throughout the session to ensure compliance with instructions. All tasks were completed using participants' own computers through a web browser. Dyads were paired based on their trait-confidence levels, which were measured in a pre-screening study. The general knowledge tests were administered in a counterbalanced order to mitigate the effects of practice and fatigue. After providing consent, participants completed the following sequence of measures: demographic questionnaire, Social Motivation Scale, general knowledge tests, Trust Scale, Psychological Safety Scale, Empathy Quotient, Reading the Mind in the Eyes, BIS BAS, and risk aversion. The order of the general knowledge tests was counterbalanced.

Participants completed each general knowledge test item twice. The first response was always given individually, with no communication permitted between dyad members. The second response, referred to as the "dyad" response, varied by the communication condition. Each trial functioned as follows. First, both participants viewed the same item and independently submitted their responses using a mouse or keyboard. After submitting an initial response, participants individually rated their confidence. For the isolated condition, participants independently repeated this response and confidence rating procedure a second time without any interaction. For the passive condition, after submitting their initial individual response and confidence rating, each participant waited for their partner to finish responding. Once both individual responses were submitted, each participant viewed their partner's response and confidence rating on screen for 5 seconds. Following this exposure, participants independently provided a second response to the same question. For the active condition, after both dyad members had submitted their individual responses and confidence ratings, participants pressed the spacebar to view the same item again. They were instructed

to discuss the question and agree upon a joint response. Responses were not displayed on their screens like in the passive condition. Following this discussion, participants submitted their joint response and then individually rated their confidence in this joint response. Then the process repeated with the next question. This process is displayed in 1.

Figure 4.1. *General Knowledge Test Procedure for Each Communication Condition*



4.2.1.5 Statistical Analyses

Unless otherwise specified, we used linear mixed-effects models (LMMs) to account for the hierarchical structure and non-independence of the data. Non-independence is inherent to repeated-measures designs, particularly when individual and dyadic responses are collected from the same participants. In our data, individuals were nested within dyads which were nested within communication conditions. All analyses were conducted at the individual level to preserve within dyad variance.

For hypotheses 1-4, we systematically compared alternative random effects structures and selected the best-fitting model based on model fit indices. Detailed model specifications and comparisons are reported in the Appendix C (see Tables C6, C11, and C16), while only the final models are reported in the main text.

To assess hypotheses 5a and 5b, we employed LPA, a person-centred statistical approach that identifies unobserved subgroups based on shared patterns on a set of psychological variables. In this study, LPA was used to classify dyads into distinct psychological profiles based on their levels of cognitive ability and trait confidence. This allowed us to investigate whether combinations of these individual-level traits were associated with different dyadic performance trajectories, similar to our prior study (Chapter 3).

All analyses were conducted using R version 4.4.2. The LMMs were estimated using the lme4 and lmerTest packages (Bates et al., 2015; Kuznetsova et al., 2017). The LPA model was estimated using the tidyLPA package (Rosenberg et al., 2018). Model selection was guided by standard fit indices, including AIC, BIC, entropy, and interpretability of the profiles. Because overall model estimates did not provide the specific comparisons required to test our hypotheses, we conducted pairwise contrasts. All *p*-values were adjusted using the Holm procedure to control the type I error rate.

4.3 Results

The results of the main study are reported here, with the results of the pre-screening study reported in Appendix C. All analyses examining differences between conditions on decision accuracy and decision confidence were conducted at the individual level. Both individual and dyadic responses were represented by each participant's average score across items, allowing comparisons across experimental conditions. This approach allowed us to

account for within-dyad variance and the nested structure of the data by modelling random effects for individuals nested within dyads and communication conditions.

4.3.1 Descriptive Statistics

4.3.1.1 Decision Accuracy and Decision Confidence

Table 4.1 presents the descriptive statistics, and internal consistency estimates for individuals and dyads for each type of communication and trait confidence condition. Internal consistency was measured using Omega total because we assumed unidimensionality but not tau-equivalence (McDonald, 1999).

Table 4.1. *Descriptive Statistics and Internal Consistency Estimates for Decision Accuracy and Decision Confidence for Each Condition (N = 210; n = 70 per Trait Confidence Condition)*

	Individuals				Dyads			
Outcome	ω_t	Low Mean (SD)	Mixed Mean (SD)	High Mean (SD)	ω_t	Low Mean (SD)	Mixed Mean (SD)	High Mean (SD)
Decision Accuracy								
Isolated	.56	72.57 (18.31)	77.57 (14.59)	75.29 (19.39)	.55	74.00 (17.89)	77.14 (15.52)	74.86 (18.08)
Passive	.64	74.26 (15.86)	76.71 (16.48)	75.14 (19.62)	.55	82.94 (11.73)	81.29 (14.64)	81.00 (16.08)
Active	.56	72.43 (17.48)	73.86 (17.55)	75.14 (17.51)	.53	79.00 (13.64)	84.14 (12.80)	85.57 (10.85)
Decision confidence								
Isolated	.82	67.72 (10.00)	71.60 (10.41)	71.97 (10.27)	.83	66.80 (10.14)	70.44 (10.59)	71.43 (10.27)
Passive	.79	68.22 (8.43)	71.38 (9.96)	71.25 (11.24)	.80	71.53 (8.03)	75.03 (9.73)	75.46 (11.10)
Active	.81	64.15 (9.24)	66.08 (9.32)	67.90 (10.21)	.84	65.79 (8.99)	68.10 (9.14)	71.88 (11.11)

Note. ω_t = Internal consistency measured using omega total.

Internal consistency estimates for decision accuracy were low (Omega total ranged from .53 to .64). While these values raise concerns about measurement precision, they were considered minimally acceptable given the exploratory nature of the research. Nonetheless, caution is warranted when interpreting the findings for decision accuracy. For decision confidence, internal consistency estimates were good (.79 to .84).

We examined the relationship between decision confidence and accuracy for individual responses to assess whether decision confidence was a reliable signal for correctness. Across all three communication conditions, the average within-individual correlation was positive, indicating that confidence generally tracked accuracy. However, the strength of this relationship differed by condition. The passive condition had a significantly stronger correlation ($r = .44$, $p < .001$, 95% CI = [.40-.48]) than both the isolated ($r = .32$, $p < .001$, 95% CI = [.28-.36]) and active ($r = .27$, $p < .001$, 95% CI = [.23-.31]) conditions, which did not significantly differ from each other. These results suggest that decision confidence was a more informative signal for decision accuracy in the passive communication condition and this difference may have influenced our findings.

4.3.1.2 Demographic and Individual Difference Measures

Table 4.2 displays the descriptive statistics and, where relevant, internal consistency estimates for the demographic, individual difference, and communication variables. The means and standard deviations are as expected given the results of other studies that have used the same measures on undergraduate populations (e.g., Blanchard et al., 2020, 2025; Jackson et al., 2017; Law et al., 2018; Law et al., 2022). Internal consistency estimates ranged from acceptable (.64) to excellent (.91) for all psychological measures.

ANOVA tests confirmed that education, cognitive ability, and trait confidence significantly differed between the three dyad types differing on trait confidence conditions. Tukey's HSD tests indicated that trait confidence and cognitive ability were significantly higher for mixed-trait vs low-trait ($p < .001$ and $p < .01$, respectively), high-trait vs low-trait ($p < .001$ for both), and high-trait vs mixed-trait ($p < .001$ for both) confidence dyads. These differences in trait confidence confirmed our manipulation check, as they were intended by design and resulted from the pre-screening study which identified high-trait (> +0.50 SD) and low-trait (< -0.50 SD) confidence individuals. Given that the present study aimed to isolate

the influence of trait confidence on dyadic outcomes, we included EAT accuracy and RAPM accuracy as covariates when examining the differences between our conditions on the general knowledge tests. Furthermore, the trait confidence conditions significantly differed on education *($F_{2,207}$ = 9.99, p < .001)*. Tukey HSD tests indicated that the high-trait confidence dyads had significantly higher levels of education than low-trait (p < .001) and mixed-trait confidence dyads (p = .03) but there was no difference between the low-trait and mixed-trait conditions (p = .14). Lastly, intellect was significantly higher for those in high-trait compared to the low-trait confidence dyads (p = .02), but the other conditions did not differ[4]. This analysis is presented in Appendix C (see Table C10). There were no significant differences between the trait confidence conditions on the other measures.

Table 4.2. *Descriptive Statistics and Internal Consistency Estimates for Demographic, Individual Difference, and Communication Variables for the Trait Confidence Conditions (N = 210; n = 70 Per Trait Confidence Condition)*

Variable	ω_t	Low Mean (SD)	Mixed Mean (SD)	High Mean (SD)	$F_{2, 207}$
Age	-	20.79 (6.40)	22.61 (6.73)	22.04 (5.33)	1.60
Education	-	1.00 (0.17)	1.30 (1.02)	1.70 (1.23)	9.99***
Proportion of females	-	0.19 (0.39)	0.30 (0.46)	0.26 (0.44)	1.25
Cognitive Ability	-	56.56 (11.12)	62.64 (13.97)	73.38 (10.78)	35.01***
RAPM accuracy	.72	50.48 (18.04)	59.71 (19.64)	75.33 (13.82)	36.74***
EAT accuracy	.64	62.64 (13.10)	65.57 (14.83)	71.43 (13.62)	7.28***
Trait Confidence	-	53.85 (7.40)	68.53 (16.62)	83.30 (5.22)	127.20***
RAPM confidence	.91	47.35 (10.57)	64.57 (19.42)	83.02 (8.17)	120.32***
EAT confidence	.91	60.35 (10.26)	72.49 (16.48)	83.58 (7.51)	65.47***
Agreeableness	.74	3.85 (0.76)	3.91 (0.69)	3.89 (0.72)	0.12
Conscientiousness	.73	3.22(0.88)	3.38(0.84)	3.40(0.86)	0.89
Extraversion	.85	3.01(0.96)	2.90(0.97)	2.65(0.98)	2.49†
Intellect	.73	3.53(0.80)	3.79(0.71)	3.88(0.69)	4.24*
Neuroticism	.71	3.15(0.77)	3.12(0.86)	3.03(0.90)	0.38

Note. ω_t = internal consistency measured using omega total.
*** p < .001, ** p < .01, * p < .05, † p < .10

[4] To ensure the difference in education did not influence our findings, we re-ran the same analyses reported in the next section with education included as a covariate. It was not a significant predictor in any of the models and the significance of the fixed effects did not change.

4.3.2 Effects of Trait Confidence and Communication on Decision Outcomes

We fit a series of LMMs to examine the effects of grouping (individual vs. dyad; within-subjects), type of communication (isolated vs. passive vs. active; within-subjects), and trait confidence (low-trait vs. mixed-trait vs. high-trait; between-subjects) on decision accuracy and decision confidence. We included two cognitive ability estimates (i.e., EAT accuracy and RAPM accuracy) as covariates to control for the effect of ability. The models included random intercepts and slopes for communication across individuals and dyads. Our interpretations focused on the three-way interaction effects. Table 4.3 presents the fixed effects and Figure 4.2 displays the differences between dyads and individuals in each condition for both outcomes.

Figure 4.2. *The Differences Between Dyads and Individuals on Decision Accuracy and Decision Confidence for All Conditions*

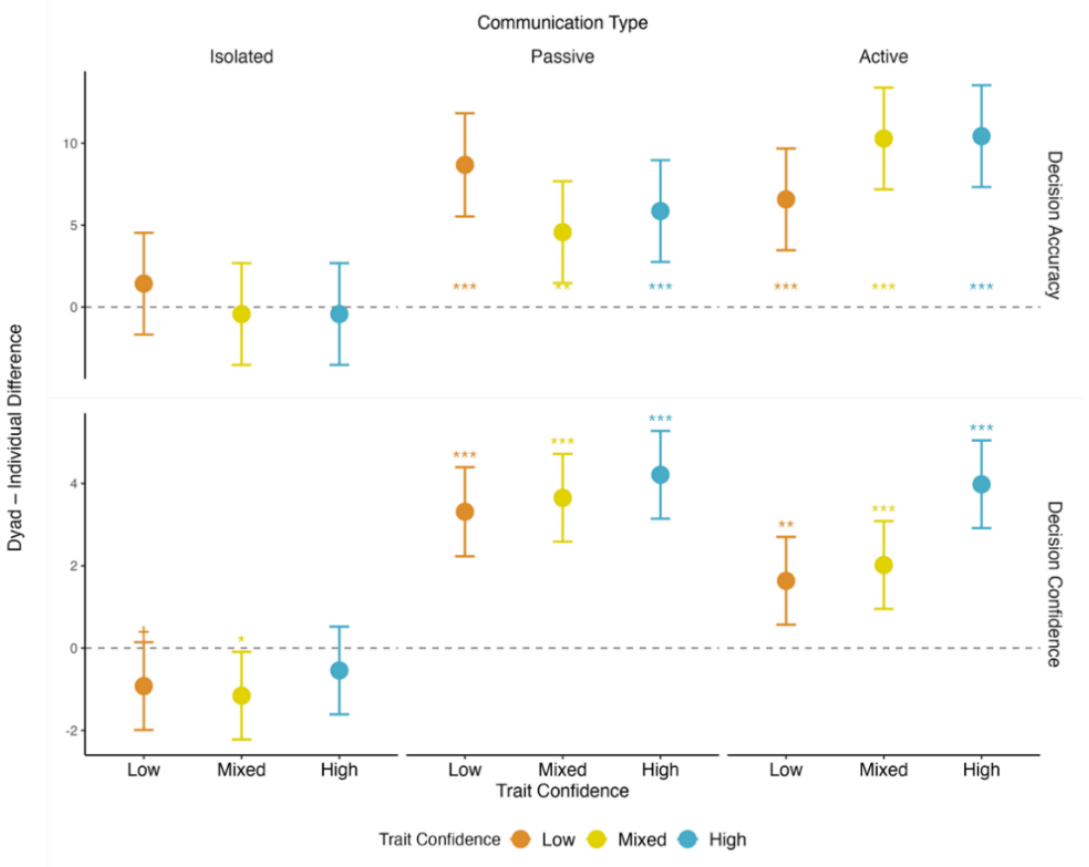


Table 4.3. *Pairwise Comparisons for Decision Accuracy and Confidence (N = 1256)*[5]

			Decision			
			Accuracy		Confidence	
			b	SE	*b*	SE
Main Effects						
Grouping	Dyad vs Ind		5.22***	0.53	1.80***	0.18
Communication	Passive vs Isolated		3.33*	1.36	2.24**	0.64
	Active vs Isolated		3.12*	1.29	-2.67***	0.57
	Passive vs Active		-0.21	1.33	-4.91***	0.61
Trait Confidence	Mixed vs Low		2.54	2.16	3.14	1.68
	High vs Low		1.67	2.32	4.47*	1.81
	High vs Mixed		-0.87	2.21	1.33	1.72
RAPM Acc			-0.07†	0.04	-0.06*	0.03
EAT Acc			0.22***	0.04	0.15***	0.03
Three-Way Simple Effects Interactions						
Trait Confidence	Communication	Grouping				
Low	Isolated	Dyad vs Ind	1.43	1.58	-0.92†	0.54
Mixed	Isolated	Dyad vs Ind	-0.43	1.58	-1.16*	0.54
High	Isolated	Dyad vs Ind	-0.43	1.58	-0.54	0.54
Low	Passive	Dyad vs Ind	8.68***	1.6	3.31***	0.55
Mixed	Passive	Dyad vs Ind	4.57**	1.58	3.65***	0.54
High	Passive	Dyad vs Ind	5.86***	1.58	4.21***	0.54
Low	Active	Dyad vs Ind	6.57***	1.58	1.63**	0.54
Mixed	Active	Dyad vs Ind	10.29***	1.58	2.02***	0.54
High	Active	Dyad vs Ind	10.43***	1.58	3.98***	0.54

*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$

4.3.2.1 Baseline Analyses

First, we examined baseline differences between the conditions by comparing individual responses between the communication and trait confidence conditions. This allowed the baseline tests and main tests to be conducted within a single model thus maintaining control over the Type I error rate. The results of all baseline comparisons are presented in Table C8 in Appendix C.

[5] $N = 210$ participants x 2 grouping conditions x 3 communication conditions – 4 missing data points = 1256. We checked if education significantly predicted any of the outcomes by refitting the LMMs with education included as a covariate. It was not a significant predictor of decision accuracy (b = -1.17, SE = 0.76, $t_{202} = -1.55$, $p = .12$) or decision confidence (b = -0.63, SE = 0.59, $t_{201} = 1.07$, $p = .29$).

Decision accuracy. No significant differences were found between the conditions on decision accuracy at baseline (individual responses).

Decision confidence. Baseline analyses showed significant differences on decision confidence between the communication conditions. Specifically, individuals in the active communication condition had lower decision confidence than those in the passive and isolated communication conditions across all three trait confidence conditions. Given that these differences existed prior to, and cannot be attributed to, the grouping intervention, we did not perform between-group simple effects contrasts comparing the communication conditions at each level of grouping or trait confidence.

4.3.2.2 Main Analyses

4.3.2.2.1 Decision Accuracy

The overall model for decision accuracy significantly differed from the null model ($\chi^2 = 176.28$, $p < .001$, $R^2 = .09$). Together, the fixed effects accounted for 9% of the variance in decision accuracy.

Three-Way Interactions. In support of hypothesis 1c and 1d, low-trait confidence dyads were significantly more accurate than individuals in the passive condition ($b = 8.68$, SE = 1.60, $t_{619} = 5.41$, $p < .001$) and the active condition ($b = 6.57$, SE = 1.58, $t_{619} = 5.41$, $p < .001$). In addition, the size of the decision accuracy improvements for low-trait confidence dyads did not differ between the active and passive communication conditions ($b = -2.11$, SE = 2.25, $t_{619} = -1.58$, $p = .11$).

For mixed-trait confidence dyads and in support of hypotheses 1a, collective judgments were significantly more accurate than individual judgments in the passive condition ($b = 4.57$, SE = 1.58, $t_{619} = 2.89$, $p < .01$) and active communication conditions

(b = 10.29, SE = 1.58, t_{619} = 6.51, p < .001). In support of hypothesis 1b, the magnitude of decision accuracy gains were significantly greater in the active compared to the passive communication condition (b = 5.71, SE = 2.24, t_{619} = *2.56*, p = .01).

In support of hypotheses 1e, high-trait confidence dyads were significantly more accurate in their collective judgments than individually for both the passive (b = 5.86, SE = 1.58, t_{619} = 3.70, p < .001) and active communication conditions (b = 10.43, SE = 1.58, t_{619} = 6.60, p < .001). In support of hypothesis 1f, the magnitude of decision accuracy improvements were significantly larger in the active than the passive communication condition (b = 4.57, SE = 2.24, t_{619} = 2.04, p = .04).

4.3.2.2.2 Decision Confidence

The overall model predicting decision confidence significantly differed from the null model (χ^2 = 277.60, p < .001, R^2 = .13), with fixed effects explaining 13% of the variance.

Three-Way Interactions. For the low-trait confidence dyads, collective judgments came with significantly higher decision confidence than individual judgments in the passive (b = 3.31, SE = 0.55, t_{619} = 6.01, p < .001) and the active communication conditions (b = 1.63, SE = 0.54, t_{619} = 3.01, p < .01). Furthermore, the increase in decision confidence (from individual to dyadic responses) was significantly greater in the passive compared to the active communication condition (b = -1.68, SE = 0.77, t_{619} = -2.17, p = .03).

For mixed-trait confidence dyads, collective judgments came with significantly higher decision confidence than those made individually in both passive (b = 3.65, SE = 0.54, t_{619} = 6.73, p < .001) and active communication conditions (b = 2.02, SE = 0.54, t_{619} = 3.72, p < .001). Similar to low-trait confidence dyads, those with mixed-trait confidence had significantly greater decision confidence gains in the passive compared to the active communication condition (b = -1.63, SE = 0.77, t_{619} = -2.13, p = .03).

High-trait confidence dyads exhibited significantly higher decision confidence for collective judgments compared to those made individually, in both the passive (b = 4.21, SE = 0.54, t_{619} = 7.76, p < .001) and active communication conditions (b = 3.98, SE = 0.54, t_{619} = 7.33, p < .001). However, unlike the low-trait and mixed-trait confidence dyads, the magnitude of their decision confidence gains did not significantly differ between the passive and active conditions (b = -0.23, SE = 0.77, t_{619} = -0.30, p = .76).

Contrary to hypothesis 2b, dyadic increases in decision confidence were greater in the passive than the active communication condition for low-trait and mixed-trait confidence dyads but there was no difference for high-trait confidence dyads.

Additionally, for the active communication condition, high-trait confidence dyads showed significantly greater decision confidence compared to both low-trait (b = 2.34, SE = 0.77, t_{619} = 3.05, p < .01) and mixed-trait confidence dyads (b = 1.96, SE = 0.77, t_{619} = 2.55, p = .01), whereas no such difference was observed between low-trait and mixed-trait confidence dyads (b = 0.38, SE = 0.77, t_{619} = 0.50, p = .62). Under the passive condition, trait confidence had no effect on decision confidence gains. That is, no differences were observed between high-trait and low-trait (b = 0.90, SE = 0.77, t_{619} = 1.16, p = .25), high-trait and mixed-trait (b = 0.56, SE = 0.77, t_{619} = 0.73, p = .47), or low-trait and mixed-trait confidence dyads (b = 0.34, SE = 0.77, t_{619} = 0.44, p = .66).

In partial support of hypothesis 2a, high-trait confidence dyads had the largest increase in decision confidence compared to low-trait and mixed-trait confidence dyads in the active condition but not the passive communication condition.

Interestingly, in the isolated condition where members answered alone both times, the second (“dyadic”) response was associated with significantly lower confidence than the first individual response in the mixed condition (b = -1.16, SE = 0.54, t_{619} = -2.13, p = .03) and

this difference approached significance in the low condition (b = -0.92, SE = 0.54, t_{619} = -1.70, p = .09).

4.3.2.3 Confidence Matching

We examined whether dyads engaged in confidence matching at the item level, a process where participants' dyadic confidence judgments became more similar to their partner's individual confidence judgements. We also explored whether the extent of confidence matching varied by communication condition or trait confidence.

To investigate, we first calculated the difference between each participant's dyadic and individual decision confidence ratings for each item, representing the change in decision confidence caused by interaction. We then aligned each participant's change in decision confidence with their teammate's individual decision confidence rating for the same item. Finally, we fit a LMM using the teammate's individual decision confidence, communication type, and trait confidence to predict the change in decision confidence while controlling for differences in two cognitive abilities. If confidence matching occurred, then a dyad member's individual decision confidence would predict their partner's change in decision confidence, particularly for the passive and active communication conditions where interaction was permitted. A higher regression coefficient for confidence matching indicated greater alignment in dyadic decision confidence ratings.

The final model included three-way interaction terms and random intercepts for group and item number. Table 4.4 presents the fixed effects relevant to confidence matching. Figure 4.3 displays the modelled slopes for confidence matching across communication and trait confidence conditions. The full model results are available in Appendix C (see Tables C13, C14, and C15).

Table 4.4. *Main Effects and Simple Interaction Contrasts for Confidence Matching (N = 6280)*[6]

Main Effect		b	SE	t
Teammate's Individual Confidence		0.15	0.01	20.30***
Interaction Effects for Teammate's Individual Decision confidence				
Communication	Trait Confidence			
Isolated	Low	0.02	0.02	1.14
Isolated	Mixed	-0.02	0.02	-0.86
Isolated	High	0.02	0.02	1.05
Passive	Low	0.28	0.02	13.46***
Passive	Mixed	0.20	0.02	10.85***
Passive	High	0.23	0.02	12.74***
Active	Low	0.19	0.02	8.54***
Active	Mixed	0.23	0.02	10.86***
Active	High	0.17	0.02	8.65***

*** $p < .001$

The overall model significantly differed from the null model ($\chi^2 = 674.58$, $p < .001$, $R^2 = .14$), accounting for 14% of the variance in confidence matching. The main effect of confidence matching was significant ($b = 0.15$, SE = .01, $t_{2811} = 20.30$, $p < .001$). Additional analyses demonstrated that the relationship between a dyad member's shift in decision confidence and their partner's individual level of decision confidence was moderated by the type of communication and trait confidence. Specifically, there were significant positive slopes in the passive (low-trait: $b = 0.28$, SE = 0.02, $t_{4673} = 13.46$, $p < .001$; mixed-trait: $b = 0.20$, SE = 0.02, $t_{4528} = 10.85$, $p < .001$; and high-trait: $b = 0.23$, SE = 0.02, $t_{5092} = 12.74$, $p < .001$) and active (low-trait: $b = 0.19$, SE = 0.02, $t_{5791} = 8.54$, $p < .001$; mixed-trait: $b = 0.23$, SE = 0.02, $t_{5942} = 10.86$, $p < .001$; high-trait: $b = 0.17$, SE = 0.02, $t_{6116} = 8.65$, $p < .001$) communication conditions, suggesting that dyad members tended to align their decision confidence in both communication conditions that permitted interaction. In contrast, this relationship did not emerge in the isolated condition (low-trait: $b = 0.02$, SE = 0.02,

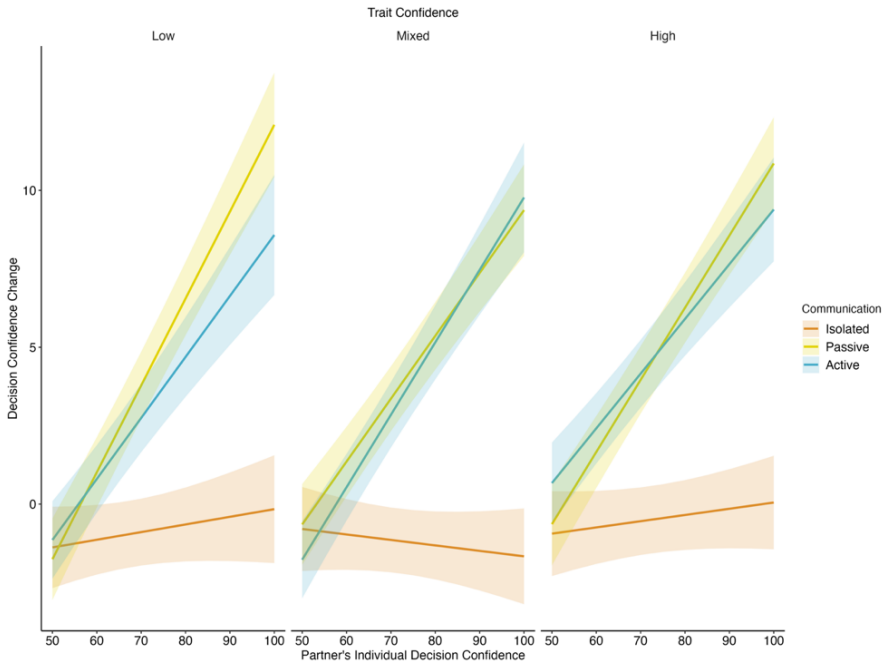
[6] 210 participants x 3 communication conditions x 10 items – 20 missing data points = 6280.

t_{6067} = 1.14, p = .26; mixed-trait: b = -0.02, SE = 0.02, t_{5751} = -0.86, p = .39; high-trait: b = 0.02, SE = 0.02, t_{5948} = 1.05, p = .30).

Furthermore, the magnitude of confidence matching slopes did not differ between the passive and active conditions for any of the trait confidence conditions (low-trait: b = 0.08, SE = 0.03, t_{5398} = 2.74, p = .10; mixed-trait: b = -0.03, SE = 0.03, t_{5447} = -1.11, p = 1.00; high-trait: b = 0.06, SE = 0.03, t_{5806} = 2.10, p = .54).

In support of hypothesis 3a, confidence matching emerged in both the passive and active conditions. However, the magnitude of the confidence matching effects did not differ between the passive and active conditions, thus, hypothesis 3b was not supported.

Figure 4.3. *Confidence Matching for the Communication and Trait Confidence Conditions*



Next, we investigated whether confidence matching predicted the change in a dyad's decision accuracy at the aggregate level. For each individual, we estimated simple slopes that captured the relationship between a dyad member's decision confidence shift and their teammate's individual decision confidence across items. We then entered the confidence matching slopes, communication type, and trait confidence (plus our two measures of

cognitive ability as covariates) into the model to predict the change in decision accuracy (difference between overall dyadic accuracy and individual accuracy). The final model included fixed effects only and two-way interactions. Table 4.5 presents the fixed effects. Full model details are provided in Appendix C (see Tables C18 and C19). Figure 4.4 illustrates the modelled slopes for the relationships between confidence matching and the decision accuracy change for the communication and trait confidence conditions.

The overall model was significant ($F_{15,600} = 7.19$, $p < .001$, $R^2 = .15$), accounting for 15% of the variance in the change in decision accuracy. The relationship between confidence matching and the change in decision accuracy was significant overall ($b = 0.06$, SE = 0.02, $t_{600} = 2.75$, $p < .01$). Further analyses showed that this relationship was moderated separately by the type of communication and trait confidence. Specifically, in the passive condition there was a significant positive relationship ($b = 0.15$, SE = 0.04, $t_{600} = 3.74$, $p < .001$), suggesting that greater confidence alignment predicted decision accuracy gains for dyads. In contrast, no relationship was observed in the isolated condition ($b = 0.01$, SE = 0.05, $t_{600} = 0.13$, $p = .90$) or the active communication condition which approached but did not reach significance ($b = 0.04$, SE = 0.02, $t_{600} = 1.69$, $p = .09$). Furthermore, the magnitude of the relationship between confidence matching and decision accuracy did not significantly differ for the passive compared to active condition ($b = 0.11$, SE = 0.05, $t_{600} = 2.28$, $p = .07$), although it approached significance. For trait confidence, there was a significant positive relationship in the low-trait confidence condition ($b = 0.09$, SE = 0.03, $t_{600} = 2.73$, $p < .01$) but not in the mixed-trait ($b = 0.07$, SE = 0.04, $t_{600} = 1.84$, $p = .07$) or high-trait confidence conditions ($b = 0.03$, SE = 0.04, $t_{600} = 0.88$, $p = .38$).

In partial support of hypothesis 4a, confidence matching positively predicted a dyad's improvement in decision accuracy under the passive but not the active communication condition. Given that the magnitude of the relationship between confidence matching and a

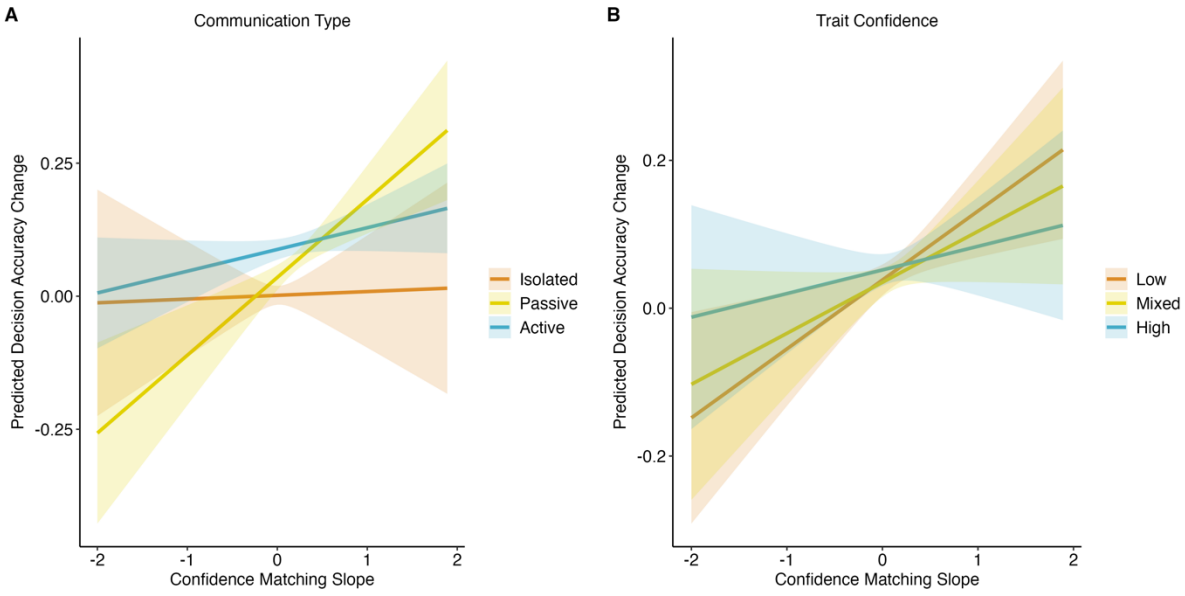
dyad's decision accuracy gains did not differ for the passive and active communication conditions, hypothesis 4b was not supported.

Table 4.5. *Main Effects and Simple Interaction Contrasts for Confidence Matching Predicting the Change in Decision Accuracy (N = 628)*

Variable	Condition	Decision Accuracy Change		
Main Effect		*b*	SE	*t*
Confidence Matching		0.06	0.02	2.75**
Interaction Effects				
Confidence Matching	Isolated	0.01	0.05	0.13
Confidence Matching	Passive	0.15	0.04	3.74***
Confidence Matching	Active	0.04	0.02	1.69†
Confidence Matching	Low	0.09	0.03	2.73**
Confidence Matching	Mixed	0.07	0.04	1.84†
Confidence Matching	High	0.03	0.04	0.88

*** $p < .001$, ** $p < .01$, † $p < .10$

Figure 4.4. *The Relationship Between Confidence Matching and the Change in Decision Accuracy for the (A) Communication and (B) Trait Confidence Conditions*



4.3.3 The Psychological Profiles of Dyads

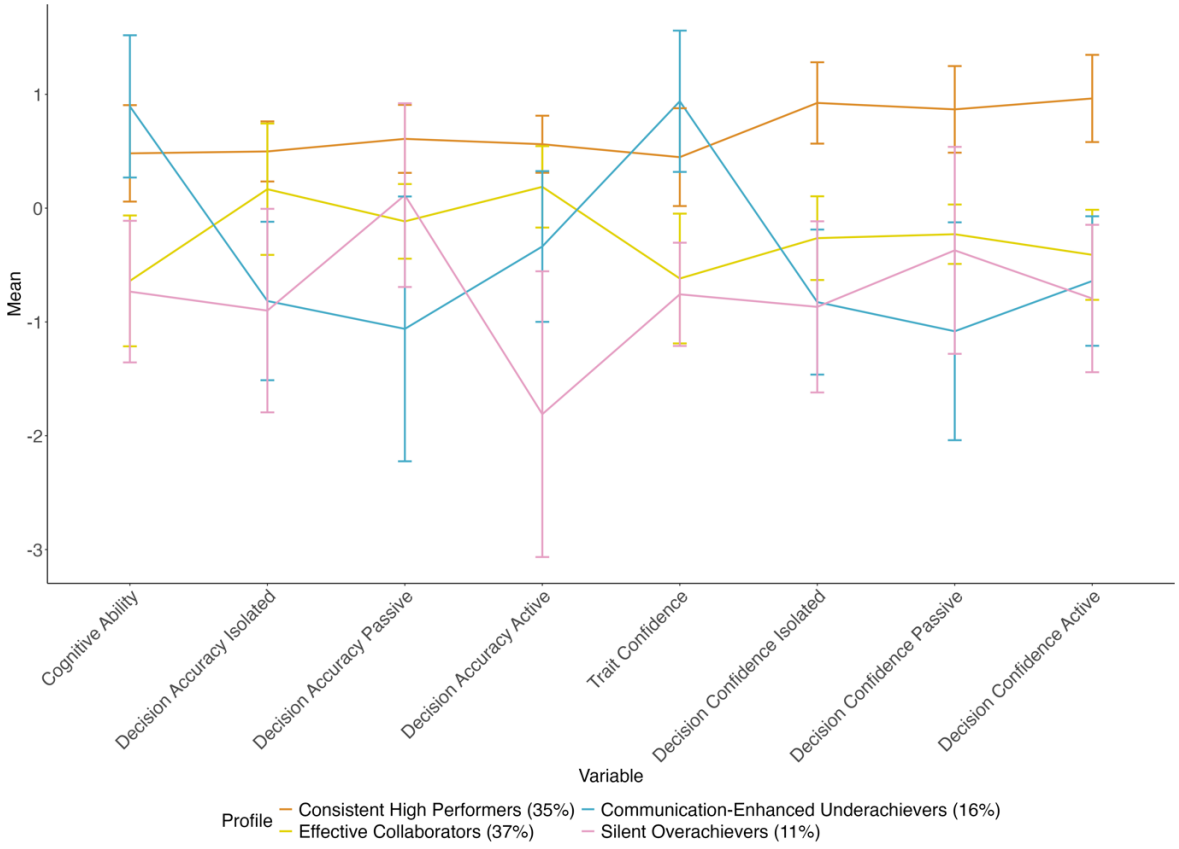
LPA does not inherently account for hierarchical data structures; therefore, we were unable to model the nesting of individuals within dyads directly. To address this, we conducted LPA at the dyad level. That is, we averaged the scores of both dyad members on

each psychological variable to produce a single set of values per dyad. The psychological profiles were then estimated via LPA using the dyad-level dataset (N = 105).

4.3.3.1 Selecting an LPA solution.

LPA was performed for solutions with 2-6 classes on eight predictor variables. These variables were cognitive ability, trait confidence, and the three dyadic measures of decision accuracy and decision confidence for the communication conditions. Scores for the isolated communication condition represented individuals because there was no interaction between dyad members. Goodness of fit statistics were used to identify the number of latent profiles (Clark & Muthén, 2009; Henson et al., 2007; Marsh et al., 2009; Spurk et al., 2020). Assessment of the indices and examination of the profiles within each model suggested a 4-Class solution was the best fitting model (See Figure 4.5). Refer to Appendix C for a detailed summary of this assessment process (see Table C20 and Figure C4).

Figure 4.5. *Mean Scores for the Four Latent Profiles*



4.3.3.2 Interpretation of the Four Profiles

Each of the four profiles had the following percentage of participants: 35.24% in Class 1 (n = 37), 37.14% in Class 2 (n = 39), 16.19% in Class 3 (n = 17), and 11.43% in Class 4 (n = 12). The profiles differed significantly on dyadic decision accuracy and decision confidence measured under the communication conditions, and cognitive ability and trait confidence. The profiles were interpreted as follows: 1) *Consistent High Performers*: Dyads with high cognitive ability and trait confidence who consistently achieved high performance regardless of the communication condition; 2) *Effective Collaborators*: Dyads with low cognitive ability and trait confidence who consistently performed at a moderate level, irrespective of the communication condition; 3) *Communication-Enhanced Underachievers*: Dyads high on cognitive ability and trait confidence who performed poorly in the isolated and passive communication conditions but achieved moderate performance under active communication; 4) *Silent Overachievers*: Dyads with low cognitive ability and trait confidence who performed poorly in the isolated and active communication conditions but achieved moderate performance under passive communication. These results support hypotheses 5a and 5b.

4.3.3.3 Differences Between the Four Profiles

First, we conducted Fisher's Exact Test on the frequency of trait confidence membership within each LPA profile. The test revealed that observed frequencies significantly differ from the expected frequencies ($p < .001$). Table 4.6 displays the observed and expected frequencies and the standardised residuals. The standardised residuals (more extreme than ±2) indicate that low trait confidence dyads were underrepresented in the Consistent High Performers and Communication-Enhanced Underachievers profiles and overrepresented in the Effective Collaborators and Silent Overachievers profiles. The

opposite was observed for high trait confidence dyads. Observed profile membership did not differ from expected membership for mixed trait confidence dyads.

Table 4.6. *Frequency Table for Membership in Trait Confidence Conditions and the LPA Profiles*

	Trait confidence		
	Low	Mixed	High
Frequencies: O(E)			
Consistent High Performers	4 (12)	13 (12)	20 (12)
Effective Collaborators	22 (13)	16 (13)	1 (13)
Communication-Enhanced Underachievers	0 (6)	4 (6)	13 (6)
Silent Overachievers	9 (4)	2 (4)	1 (4)
Std. Residuals			
Consistent High Performers	-3.61	0.29	3.32
Effective Collaborators	3.86	1.29	-5.14
Communication-Enhanced Underachievers	-3.18	-0.94	4.12
Silent Overachievers	3.25	-1.30	-1.95

Note. O = observed frequency. E = expected frequency.

We conducted a series of univariate ANOVAs to identify differences between the profiles on the individual difference variables not included in the LPA model. These variables included: social sensitivity, inequality of communication, total communication, proportion of females, big-five personality traits, psychological safety, trust, empathy, BIS/BAS scales, social motivation traits (i.e., proself, prosocial, and fearful), and risk aversion. These constructs are described in detail in Appendix C. See Table 4.7 and Figure 4.6 for a summary of the results of these analyses.

The series of ANOVAs revealed that the four LPA profiles significantly differed on the inequality of words spoken ($F_{3,100} = 4.33$, $p < .01$, $\eta_p^2 = .11$), total number of speaking turns ($F_{3,100} = 3.73$, $p = .01$, $\eta_p^2 = .10$), proportion of females ($F_{3,101} = 2.79$, $p = .04$, $\eta_p^2 = .08$), disagreement, ($F_{3,101} = 6.26$, $p < .001$, $\eta_p^2 = .16$), and extraversion ($F_{3,101} = 4.01$, $p < .01$, $\eta_p^2 = .11$). The differences including the effect sizes were small, with partial eta squared values ranging from .08 to .16. Tukey-Kramer post-hoc tests were conducted on the

significant outcomes to test differences between the profiles. Post-hoc tests indicated that Consistent High Performers (M = 78.39, SD = 57.07) had significantly lower inequality of words spoken than Effective Collaborators (M = 148.62, SD = 113.51, p < .01). Furthermore, Consistent High Performers (M = 99.67, SD = 22.98) had significantly fewer speaking turns than Communication-Enhanced Underachievers (M = 125.59, SD = 37.14). Additionally, Consistent High Performers (M = 28.38, SD = 15.90) had significantly less disagreement than Effective Collaborators (M = 43.59, SD = 17.24, p < .001). Lastly, Communication Enhanced Underachievers (M = 9.38, SD = 2.32) had significantly lower extraversion than Consistent High Performers (M = 11.55, SD = 3.02, p = .04), Effective Collaborators (M = 11.86, SD = 2.67, p = .01), and Silent Underachievers (M = 12.38, SD = 2.47, p = .02). After controlling the type I error rate, there were no differences between the profiles on the proportion of females.

Figure 4.6. *Differences Between the Four Latent Profiles*

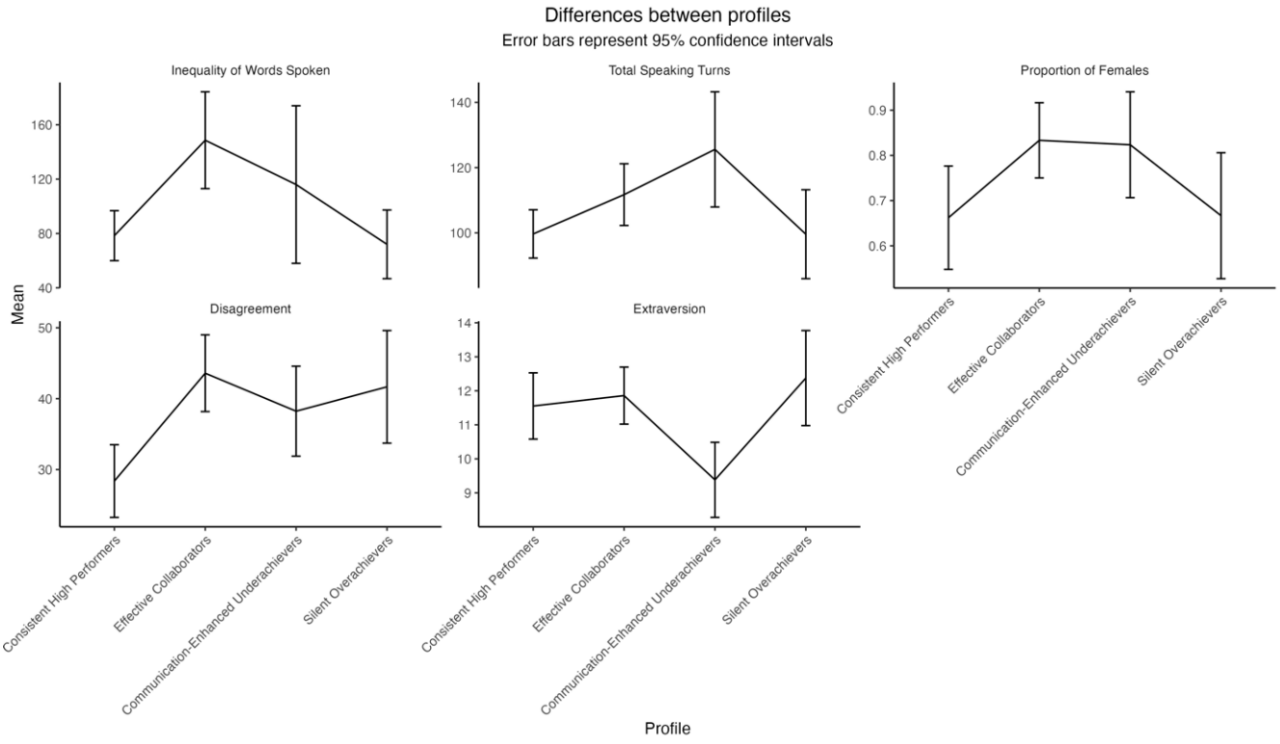


Table 4.7. *Results of ANOVAs for the Differences Between LPA Profiles (N = 105)*

Measure	Consistent High Performers Mean (SD)	Effective Collaborators Mean (SD)	Communication-Enhanced Underachievers Mean (SD)	Silent Overachievers Mean (SD)	$F_{3,101}$	η_p^2	Contrast tests c1-2	c1-3	c1-4	c2-3	c2-4	c3-4
Woolley et al. (2010) variables												
Social sensitivity	79.73 (17.52)	73.93 (16.35)	73.53 (14.8)	73.61 (18.41)	1.01	.03	.44	.59	.69	1.00	1.00	1.00
Number of females	0.66 (0.35)	0.83 (0.26)	0.82 (0.25)	0.67 (0.25)	**2.79***	.08	.06†	.25	1.00	1.00	.33	.50
Inequality of turn taking	2.47 (1.93)	2.59 (2.24)	3.62 (4.97)	1.71 (1.30)	1.26	.04	1.00	.48	.83	.56	.76	.25
Other variables												
Inequality of Words	78.39 (57.07)	148.62 (113.51)	116.01 (121.96)	72.01 (44.65)	**4.33****	**.11**	**<.01****	.52	1.00	.62	.07†	.59
Total Talking Turns	99.67 (22.98)	111.72 (30.17)	125.59 (37.14)	99.58 (24.09)	**3.73***	**.10**	.27	.01*	1.00	.34	.57	.08†
Total Words Spoken	727.31 (284.63)	905.15 (381.36)	940.18 (402.01)	732 (294.81)	2.58†	.07	.12	.16	1.00	.99	.43	.38
Disagreement	28.38 (15.90)	43.59 (17.24)	38.24 (13.34)	41.67 (14.03)	**6.26*****	**.16**	**<.001*****	.15	.06†	.65	.98	.94
Agreeableness	15.54 (2.17)	15.64 (1.96)	15.38 (1.76)	15.46 (2.55)	0.07	.00	1.00	.99	1.00	.97	.99	1.00
Conscientious	13.34 (2.08)	13.26 (2.77)	13.65 (3.08)	13.12 (2.31)	0.12	.00	1.00	.98	.99	.95	1.00	.95
Extraversion	11.55 (3.02)	11.86 (2.67)	9.38 (2.32)	12.38 (2.47)	**4.01****	.11	.96	**.04***	.80	**.01***	.94	**.02***
Intellect	14.85 (2.79)	14.92 (2.51)	15.41 (2.08)	14.58 (1.87)	0.30	.01	1.00	.87	.99	.91	.98	.81
Neuroticism	12.24 (2.52)	12.60 (2.27)	12.06 (2.72)	12.71 (2.24)	0.32	.01	.92	.99	.94	.87	1.00	.89
Psychological Safety	5.77 (0.51)	5.66 (0.56)	5.56 (0.51)	5.42 (0.69)	1.40	.04	.85	.57	.24	.91	.54	.91
Trust	4.11 (0.35)	4.14 (0.35)	4.04 (0.37)	4.22 (0.46)	0.57	.02	.99	.92	.82	.81	.91	.59
Empathy	41.38 (8.76)	44.73 (7.59)	44.97 (6.8)	45.08 (9.34)	1.47	.04	.28	.44	.52	1.00	1.00	1.00
BIS Total	21.78 (2.53)	22.29 (2.51)	21.68 (1.92)	21.79 (2.73)	0.40	.01	.80	1.00	1.00	.82	.93	1.00
BAS Drive	10.77 (1.59)	10.99 (1.73)	10.88 (1.42)	11.42 (1.68)	0.49	.01	.94	1.00	.63	1.00	.85	.82
BAS Fun	12.38 (1.54)	11.73 (1.63)	12.00 (1.41)	12.08 (1.65)	1.09	.03	.28	.84	.94	.93	.90	1.00
BAS Reward	17.07 (1.55)	17.24 (1.44)	17.12 (1.50)	17.75 (1.06)	0.70	.02	.95	1.00	.49	.99	.72	.66
Proself Motivation	-0.02 (0.31)	-0.01 (0.28)	-0.04 (0.33)	0.13 (0.44)	0.75	.02	1.00	1.00	.53	.98	.60	.52
Prosocial Motivation	0.02 (0.33)	0.03 (0.35)	0.00 (0.27)	-0.15 (0.50)	0.79	.02	1.00	1.00	.49	.99	.46	.71
Fearful Motivation	-0.05 (0.36)	0.03 (0.36)	-0.03 (0.39)	0.10 (0.41)	0.66	.02	0.74	1.00	.62	.94	.96	.81
Risk Aversion	4.49 (1.28)	4.59 (1.82)	4.59 (1.23)	4.62 (1.79)	0.04	.00	0.99	1.00	.99	1.00	1.00	1.00

Note. η_p^2 = partial eta squared; c = contrast. Significant F values, relevant eta squared values, and contrast tests are in bold.

*** $p < .001$, ** $p < .01$, * $p < .05$, $^{\dagger}p < .10$

4.4 Discussion

In the present study, we examined how trait confidence and the type of communication influenced decision accuracy, decision confidence, and confidence matching in dyads. By selecting participants with similar cognitive abilities and statistically controlling for it, we ensured that the observed effects primarily reflected differences in trait confidence rather than ability. This study offers several novel contributions: 1) we demonstrated that trait confidence moderates dyadic gains in both decision accuracy and confidence; 2) we observed that these moderating effects depend on the mode of communication (passive vs active); 3) we showed that confidence matching naturally emerges when dyads engage in realistic verbal interactions; and 4) we identified distinct psychological profiles of dyadic performance.

4.4.1 Decision Accuracy: Interaction of Communication and Trait Confidence

Consistent with prior research, we observed a "two heads are better than one" effect when dyads were allowed to communicate (both passive and active; e.g., Bahrami et al., 2010; 2012a; Koriat, 2015). Importantly, we extended previous findings by showing that the size of this dyadic advantage depended on both the dyad's trait confidence composition and the type of communication. Specifically, mixed-trait and high-trait confidence dyads demonstrated larger accuracy improvements in the active communication condition, whereas low-trait confidence dyads benefitted equally from passive and active communication. These effects are unlikely to reflect differences in cognitive ability, as cognitive ability was constrained during recruitment and statistically controlled for in our analyses. We also observed no baseline accuracy differences between trait confidence conditions and found comparable overall gains across trait confidence conditions, though these occurred under different communication conditions.

For mixed-trait confidence dyads, these findings support the confidence matching theory (Bang et al., 2017) which suggests that large initial discrepancies in decision confidence can undermine collective performance. Under passive communication, where members can view each other's answers but cannot clarify or negotiate, such discrepancies may limit information integration, especially if the high-trait confidence member dominates decisions. In contrast, active verbal communication may have reduced this asymmetry, likely by compelling high-trait confidence members to engage more openly with their partner's perspective. This aligns with Koriat's (2024) proposal that confidence reflects response replicability: high-trait confidence individuals may rigidly adhere to initial judgments unless prompted to revise them through richer social interaction and consensus.

Dyads composed of similarly confident members showed distinct patterns. Low-trait confidence dyads achieved consistent accuracy gains across both communication types, suggesting a general openness to external input and greater willingness to revise initial judgments (Koriat, 2024). By contrast, high-trait confidence dyads, showed larger accuracy improvements under active verbal discussion, likely because verbal interaction forced reconsideration of their judgments. Under passive conditions, high-trait dyads may have discounted their partner's input, relying more heavily on their own judgments.

Together, our findings reinforce the central claim of the confidence theory: that communicating subjective decision confidence enhances dyadic performance (Bahrami et al., 2010). More importantly, our results highlight that trait confidence shapes how communication influences joint accuracy. Verbal communication appears more important when a dyad contains at least one high-trait confidence member. Considering trait confidence could help improve dyadic decision-making outcomes.

4.4.2 Decision Confidence: Passive Gains and Active Uncertainty

We replicated a well-established finding that dyads have higher-decision confidence than individuals (e.g., Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997). Importantly, this increase in confidence was not uniform across dyads. Consistent with our previous finding (Blanchard et al., 2020), we found that trait confidence moderated dyadic confidence gains but only under active communication. Specifically, dyads composed of high-trait confidence members showed the largest increases in decision confidence during verbal discussion. This contrasts with Schuldt et al. (2017), who reported greater confidence gains for low-trait confidence dyads. This discrepancy likely stems from methodological differences in measuring trait confidence. Our approach, like Blanchard et al. (2020), employed multiple validated cognitive tests that reliably captured stable, domain-general individual differences in trait confidence. Alternatively, Schuldt and colleagues used a single matched version of their general knowledge test with no correlation between decision accuracy and confidence. Their measure may not have been tapping into the confidence trait.

Our findings suggest a nuanced relationship between communication, trait confidence, and dyadic confidence gains. Low-trait and mixed-trait confidence dyads experienced larger increases in decision confidence in the passive condition, whereas high-trait confidence dyads showed greater increases in the active condition. This pattern suggests that verbal discussion makes uncertainty more salient, as it requires dyad members to consider alternative viewpoints and openly debate (Tindale & Sheffey, 2002). Low-trait confidence members, who are generally less decisive (Jackson et al., 2017), may be especially sensitive to this explicit uncertainty, potentially dampening their decision confidence in the active communication condition.

This interpretation aligns with prior work by Pescetelli and Yeung (2020), who found that decision confidence increases following initial agreement and decreases after

disagreement. In our task, similar to prior studies, dyads agreed on a large majority of trials (Koriat, 2015 reported 82% agreement; our sample was similar), so high levels of agreement likely amplified overall decision confidence. However, active communication may highlight the presence of disagreement more clearly than passive communication, especially for low-trait confidence individuals. We speculate that trait confidence may moderate how dyads respond to agreement and disagreement at the item-level under active communication. Our results suggest that high-trait confidence individuals may increase more in decision confidence when agreement occurs, and low-trait confidence members may decrease in decision confidence more when disagreement occurs.

Two key insights emerge from our findings. First, passive communication may boost dyadic confidence more than active discussion by obscuring awareness of disagreement. Second, trait confidence moderates how dyads respond to collaboration, shaping both the magnitude and direction of confidence gains. These findings refine our understanding of when and why dyadic confidence is amplified and highlight the importance of considering individual differences in trait confidence in models of collaborative decision making.

4.4.3 Confidence Matching: Communication Mode Matters

We found evidence that confidence matching naturally occurs under both passive and active communication, and it occurred independently of trait confidence. Crucially, confidence matching predicted decision accuracy improvements in the passive but not the active communication condition. Several factors may have contributed to this difference.

First, the nature of communication differed distinctly between the conditions. Passive communication explicitly provided numeric confidence values, offering a precise heuristic for identifying correct responses. Active verbal communication involved richer discussion but more ambiguous communication of decision confidence, reducing reliance on precise

numerical confidence ratings (Fusaroli et al., 2012). Thus, despite rapid confidence alignment, numeric based confidence matching in the passive condition may have had greater predictive power due to its precision and clarity.

Second, the baseline correlation between decision confidence and accuracy was significantly stronger in the passive (r = .42) compared to the active communication condition (r = .28). This difference implies that decision confidence provided a more reliable accuracy signal in the passive condition, possibly making confidence matching a more effective predictor of decision accuracy gains. Future research could test whether stronger baseline confidence-accuracy correlations amplify the effectiveness of confidence matching.

Third, our brief task and lack of feedback likely limited the strength of confidence matching effects, particularly under more complex verbal interactions. Longer tasks, like those used by Bang et al. (2017) and Pescetelli & Yeung (2022), might yield stronger effects, especially as dyads tend to develop a shared mostly non-numeric language for the expression of confidence when communicating verbally (Fusaroli et al., 2012). Notably, the relationship between confidence matching and decision accuracy gains approached significance under active verbal interaction, suggesting that a clearer effect may emerge under different task conditions.

Interestingly, although confidence matching predicted decision accuracy gains in the passive condition, dyads in the active condition demonstrated larger overall accuracy improvements. This suggests that mechanisms beyond confidence alignment played substantial roles in the active condition. For example, argument quality has been shown to overcome the influence of confidence in collective decisions (Trouche et al., 2014). These are important considerations for future studies.

4.4.4 Distinct Psychological Profiles of Dyads

Our LPA identified four distinct profiles describing dyadic behaviour across communication conditions, largely corresponding with trait confidence levels. High-trait confidence dyads typically formed either Consistent High Performers (~35% of all dyads) or Communication-Enhanced Underachievers (~16% of all dyads), whereas low-trait confidence dyads were mostly classified as Effective Collaborators (~37% of all dyads) or Silent Overachievers (~11% of all dyads). Mixed-trait confidence dyads were relatively evenly distributed among the four profiles.

These findings reinforce that trait confidence can inform the formation of dyads based on communication conditions. Specifically, high-trait confidence dyads maintain or improve performance most effectively under active communication, whereas low-trait confidence dyads perform equally well or better under passive communication. While we limited variability in cognitive ability to isolate the effects of trait confidence, prior research has demonstrated the importance of cognitive similarity and dyadic accuracy (e.g., Bahrami et al., 2010), suggesting that dyad composition should consider both cognitive ability and trait confidence.

The four profiles also differed on several psychological and coordination constructs tested outside the LPA model. Consistent High Performers, who excelled under active communication, exhibited lower inequality of words spoken and fewer disagreements than Effective Collaborators, though these differences did not extend to the lower performing profiles. Similarly, Consistent High Performers used fewer speaking turns than Communication-Enhanced Underachievers under active communication, indicating that focused and efficient interactions may underpin their strong performance. Furthermore, Communication-Enhanced Underachievers displayed lower extraversion scores than the other

profiles, suggesting that limited sociability or assertiveness may impede their performance, requiring active communication to partially mitigate this deficit.

Interestingly and contrary to prior research linking collective intelligence to social sensitivity, equality of turn-taking, and the proportion of females (e.g., Aggarwal et al., 2019; Engel et al., 2014, 2015; Woolley et al., 2010), we did not observe systematic differences on these variables between the profiles. This indicates that trait confidence, cognitive ability, and efficient communication styles may outweigh these factors in driving dyadic performance.

Overall, the LPA results complement our LMM findings, confirming that both trait confidence and the communication type shape how dyads coordinate, negotiate, and perform collaboratively. High-trait confidence dyads consistently excel, likely influenced by efficient communication and lower levels of disagreement, while low-trait confidence dyads benefit from simpler, passive communication conditions. These findings highlight the combined importance of individual attributes (i.e., trait confidence, cognitive ability, and personality) and group-level communication dynamics in determining distinct trajectories of dyadic performance.

4.4.5 Implications and Contributions to Theory

Our findings have important theoretical implications for the confidence-matching literature. For example, Pescetelli and Yeung (2022) raised the “similarity hypothesis” which states that observed confidence matching could reflect pre-existing similarities rather than interaction driving convergence. Our results with mixed-trait confidence dyads provide direct evidence that confidence matching naturally emerges due to social interaction rather than pre-existing similarity. Despite these dyads comprising members with substantially different baseline levels of trait confidence, confidence matching occurred only in conditions that permitted communication.

Our findings also offer insights into the “two heads are better than one” effect by demonstrating that trait confidence significantly moderates dyadic accuracy benefits and decision confidence gains through communication. This suggests trait confidence may influence how effectively dyads collaborate, share information, and possibly calibrate their confidence judgments during collective tasks.

Our findings also hold important implications for human-AI collaboration. Recent research indicates that large language models display overconfidence when faced with novel or challenging problems that are not well represented in their training data (Phan et al., 2025). This tendency could limit their effectiveness in collaborative decision-making contexts where decision accuracy is critical. However, research also shows that hybrid collectives combining human and AI inputs typically outperform both human-only and AI-only groups (Zöller et al., 2024). Our study contributes to this emerging literature by suggesting that individual differences in human trait confidence might further optimise such hybrid collaborations. Specifically, pairing humans with complementary levels of trait confidence with AI systems may mitigate the AI’s overconfidence by introducing beneficial uncertainty, thus promoting more accurate and calibrated joint decisions. Future research should explore whether human trait confidence influences the outcomes of human-AI collaboration.

4.4.6 Practical Implications

Our findings have several practical implications for organizations, managers, and groups. For high-trait and mixed-trait confidence dyads, active verbal communication maximized dyadic accuracy, whereas low-trait confidence dyads performed equally well with simpler passive communication. Given that passive communication requires fewer organizational resources (i.e., shorter meetings, reduced logistical overhead, and lower financial costs) and fewer individual resources (i.e., less time spent on task and lower cognitive load), passive communication may be ideal for low-trait confidence dyads.

Conversely, high-trait and mixed-trait confidence dyads face a resource-accuracy trade-off, with active communication enhancing decision accuracy but requiring additional time and effort. Managers could optimize dyadic performance and resource allocation by strategically pairing individuals according to trait confidence and selecting suitable communication methods. Passive communication may be sufficient for high-trait or low-trait confidence dyads when maximizing decision accuracy is less critical, but active verbal discussion is necessary for mixed-trait confidence dyads to reliably achieve the benefits of collective decision making.

4.4.7 Limitations and Future Directions

Several limitations highlight areas of focus for future research. First, although our general knowledge tests provided valuable insights, their internal consistency for decision accuracy was low (.53-.64) which suggests that caution is warranted when interpreting the associated findings. The heterogenous range of topics (e.g., art, history, geography, science, film) may have contributed to the lower reliability. Future studies should employ validated measures of general knowledge to enhance reliability and replicate our findings.

Second, our tests contained only 10 items each. Although confidence matching emerges rapidly (Pescetelli and Yeung, 2022), the brevity of our tests may have constrained its relationship with decision accuracy, especially under active communication. Longer tasks should allow dyad members to better calibrate their confidence scales.

Third, to disentangle the relationship between trait confidence and cognitive ability, we restricted variability in cognitive ability to approximately ±1.50 SD of the mean. Although this enabled clearer conclusions about trait confidence, it limits generalisability. Dyads with more pronounced ability differences might experience outcomes that differ from our findings, particularly given that substantial ability differences can undermine dyadic

benefits (e.g., Bahrami et al., 2010; Bang et al., 2017). Future studies should explore how cognitive ability differences interact with trait confidence in dyadic decision-making.

Lastly, our tests involved non-misleading items where the confidence-accuracy correlation is typically positive. An interesting direction for future research would be to examine whether confidence matching emerges for misleading items, where confidence negatively relates to decision accuracy. This could help to further delineate the boundary conditions for confidence matching as an effective heuristic strategy.

4.4.8 Conclusion

Our findings extend our understanding of dyadic decision-making by characterising how individual differences in trait confidence and communication jointly shape dyadic decision accuracy, decision confidence, and confidence matching. We found evidence that trait confidence moderates both decision accuracy and confidence gains in dyadic decision making, especially under conditions enabling active verbal communication. We also demonstrated that confidence matching occurs naturally when verbal interaction is permitted, extending prior research beyond artificial numeric confidence conditions into more realistic verbal communication contexts. Critically, we found that confidence matching predicted improvements in decision accuracy under passive communication when only numeric confidence ratings were shared. In contrast, richer verbal interactions with more ambiguous expressions of confidence led to even greater decision accuracy gains, likely through other mechanisms. Our findings highlight the relationship between confidence matching and decision accuracy as nuanced and likely contingent upon the type of communication. By demonstrating the importance of trait confidence and communication mode, our research may inform future recommendations for enhancing dyadic decisions.

Chapter 5: General Discussion

5.1 General Overview

This thesis reports on a series of three studies investigating how individual differences in cognitive ability, trait confidence, and communication shape the 2HBT1 effect and dyadic decision confidence across dynamic and static task conditions. The studies were designed to move from broader environmental influences (i.e., task conditions) toward a more fine-grained examination of psychological factors (e.g., trait confidence). Together they form a cohesive investigation into the boundary conditions of the 2HBT1 effect and the psychological underpinnings of dyadic decision-making.

Study 1 served as the starting point. It examined the 2HBT1 effect in a dynamic decision environment where task conditions change rapidly and information is asymmetrically distributed between dyad members. This study also investigated how communication quality related to the 2HBT1 effect (Chapter 2). Study 2 shifted the focus to examine the 2HBT1 effect for static, well-structured tasks. It provided a controlled context to investigate the influences of cognitive ability, metacognitive confidence, and communication on dyadic collective intelligence (Chapter 3). Study 3 extended these findings by investigating moderators of the 2HBT1 effect for static, well-structured tasks. It examined how trait confidence and communication modality shape the 2HBT1 effect, dyadic decision confidence, and confidence matching (Chapter 4).

This chapter synthesises the findings across the three studies, highlighting their theoretical and empirical contributions to the literature on the 2HBT1 effect, collective intelligence, and dyadic decision-making. It also discusses the strengths and limitations of our approach, considers implications, and outlines directions for future research.

5.2 Main Contributions to the Existing Literature

This thesis makes several novel contributions to the existing literature on dyadic decision-making. The three empirical studies highlight the roles of cognitive ability, trait confidence, and communication in shaping effective collaboration under different task conditions.

5.2.1 Cognitive Ability and Effective Collaboration

While cognitive ability was a secondary factor (covariate) in Study 1 and Study 3, it played a critical role in dyadic outcomes, although its influence was bounded by task characteristics. The main findings of each study are summarised below.

In Study 1 (Chapter 2), we observed that cognitive ability, referred to as the "competence factor," did not predict accuracy or speed on a dynamic real-time driving simulation requiring adaptation to changing conditions and communication under time pressure. This finding is not surprising given that driving engages a broad range of abilities, including visual perception, motor skills, emotional control, information processing, and executive functions (Anstey, Wood, Lord, & Walker, 2005; Asimakopulos et al., 2012, Mathias & Lucas, 2009).

In contrast, Study 2 (Chapter 3), which employed static and well-structured tasks, showed that cognitive ability was central to collective intelligence. This finding directly challenges Woolley et al.'s (2010) claim that collective intelligence is primarily shaped by social rather than cognitive factors. LPA results demonstrated that a dyad's level of cognitive ability had implications for collective performance: high-ability dyads were consistently superior, moderate-ability dyads benefitted more than expected from collaboration; and low-ability dyads consistently performed poorly.

Study 3 (Chapter 4) which also employed static and well-structured tasks, supported these patterns, even with constrained cognitive ability variance. LPA again identified profiles defined by cognitive ability, but with different trajectories depending on communication conditions. Two profiles, defined by higher cognitive ability, displayed markedly different trajectories across communication conditions. Consistent High Performers achieved constant high performance under both passive and active communication. Whereas Communication-Enhanced Underachievers performed poorly in passive communication and moderately under active communication. Similarly, two profiles characterised by lower cognitive ability showed divergent trajectories: Silent Overachievers performed better than expected under passive communication but poorly under active communication, whereas Effective Collaborators performed moderately well under all conditions but did not exceed individual performance.

Overall, these findings highlight boundary conditions for cognitive ability's role in dyadic outcomes. Cognitive ability is central to collective success in static, well-structured tasks but loses predictive utility in dynamic, continuous-performance tasks where motor skills, visual perception, and communication quality become more important. Furthermore, higher cognitive ability alone does not guarantee superior dyadic performance; other factors such as communication modality and confidence alignment also play critical roles.

5.2.2 Trait Confidence as a Moderator

Trait confidence consistently emerged as a significant predictor of accuracy improvements, escalation of decision confidence, and confidence matching, although its influence was shaped by task characteristics. The main findings of each study are summarised below.

In Study 1 (Chapter 2), we assessed confidence using two methods: cognitive confidence (using a cognitive ability task similar to our measures of trait confidence in the other two studies) and simulation confidence (post-task self-assessment). Neither measure predicted dyadic performance in the dynamic driving simulation. This likely reflects the nature of dynamic tasks like driving simulations, where participants must make hundreds of rapid decisions under time pressure, rendering a single post-task confidence rating too crude to capture the complexity of real-time judgment processes.

By contrast, in the well-structured static tasks used in Studies 2 and 3, trait confidence consistently predicted collective intelligence and dyadic accuracy gains. Study 2 (Chapter 3) showed that dyads with higher trait confidence produced more accurate decisions. Study 3 (Chapter 4) extended these findings by offering a more detailed analysis of the effect of trait confidence on decision accuracy and confidence. First, trait confidence moderated dyadic accuracy gains under different types of communication: low-trait confidence dyads benefited equally from passive and active communication, whereas mixed-trait and high-trait confidence dyads gained more from active communication. Second, trait confidence moderated decision confidence escalation, with high-trait confidence dyads increasing the most in decision confidence when working together. Third, confidence matching occurred naturally in dyads, and this effect was strongest for low-trait confidence dyads. Finally, confidence matching predicted overall accuracy gains under passive but not active communication. Trait confidence also differentiated the latent profiles identified in Studies 2 and 3. Profiles defined by higher cognitive ability generally displayed higher levels of trait confidence, whereas profiles with lower cognitive ability tended to have lower trait confidence. Together, these findings suggest that trait confidence plays a critical role in static, well-structured decision-making tasks, but has little predictive value in dynamic, real-time

environments where rapid decision-making and communication quality dominate performance.

5.2.3 Communication and its Interaction with Trait Confidence

Across all studies, communication emerged as a central determinant of collaborative performance, although the nature of its effects varied depending on task structure and individual differences in trait confidence.

Study 1 (Chapter 2) focused explicitly on the quality and quantity of communication between dyad members. Using a novel method that captured the timing and accuracy of communication exchanges, we showed that under stable (normal) operational conditions, dyads benefited from lower levels of low-quality communication, achieving higher accuracy than individuals, but they did not differ on speed. However, when operational conditions shifted unexpectedly (fog condition), dyads struggled to exploit their informational advantages. Degraded communication quality and cognitive overload led to the erosion of accuracy gains and slower speeds compared to individuals. Importantly, communication quality emerged as a stronger predictor of accuracy, while communication volume was a stronger predictor of speed. These findings illustrate that the different dimensions of communication, quality and volume, make distinct contributions to the 2HBT1 effect in dynamic, real-time tasks.

Study 2 (Chapter 3) challenged Woolley et al.'s (2010) finding that equality of participation is universally associated with collective intelligence. Consistent with independent replications (e.g., Barlow & Dennis, 2016; Bates & Gupta, 2017; Rowe et al., 2024), equality of turn taking did not predict collective intelligence overall. However, the LPA results showed that the Amplified Collective Intelligence profile, characterised by greater collective than individual performance, benefited from strategic inequality of turn

taking. In these dyads, the more capable member dominated decision making. This finding suggests that: 1) the benefits of equal participation depend on the specific context and individual differences; and 2) other communication characteristics, such as the quality or modality, may be more critical for enhancing dyadic outcomes.

Study 3 (Chapter 4) extended these findings by showing that the effects of communication on dyadic performance and confidence alignment depended on trait confidence and communication modality. Low-trait confidence dyads performed equally well under passive and active communication, while mixed-trait and high-trait confidence dyads performed best under active communication. The LPA profiles in Study 3 portrayed a more nuanced view. Equality of turn taking did not predict overall dyadic accuracy or directly distinguish profiles. However, Consistent High Performers and Silent Overachievers, displayed greater equality in speech (words spoken). Interestingly, they showed distinctly different outcomes: Consistent High Performers were defined by high-trait confidence and maintained high performance across communication conditions, whereas Silent Overachievers were characterised by low-trait confidence and had amplified performance under passive communication but poor performance under active communication.

Overall, these findings demonstrate that communication equality is not universally beneficial. Rather, it is context-dependent and the most effective communication strategies depend on the interplay between individual differences and task characteristics. Across all three studies, communication quality and communication mode consistently emerged as critical levers for dyadic effectiveness, with dynamic tasks placing greater demands on communication quality, and well-structured static tasks favouring strategic and adaptive, rather than purely egalitarian, communication patterns.

5.2.4 The Psychological Profiles of Dyads

Distinct psychological profiles of dyadic performance emerged in both Study 2 (Chapter 3) and Study 3 (Chapter 4), with considerable consistency despite differences in design. Study 3 extended the model developed in Study 2, providing a more nuanced view of how cognitive ability, trait confidence, and communication mode shape dyadic outcomes.

A key addition in Study 3 was the inclusion of the isolated communication condition, which captured individual baseline performance on the same task used to assess dyadic performance. This allowed for direct comparisons between individual performance and dyadic performance, as well as alignment with cognitive ability. The profiles demonstrated that cognitive ability did not always align with baseline individual performance: two profiles followed expected patterns (one with high and one with low ability), but the other two showed substantial mismatch, either underperforming relative to high cognitive ability or overperforming relative to low cognitive ability.

Focusing on performance in the active communication condition, both studies identified recurring profile types: consistent high performers, consistent low performers, and amplified performers. In Study 2, the amplified profile was defined by moderate levels of cognitive ability and trait confidence. In Study 3, the Communication-Enhanced Underachievers profile captured amplified performance but was composed of dyads with high cognitive ability and trait confidence. This shift may reflect key design differences in Study 3, including the pairing of participants based on trait confidence, constrained variance in cognitive ability, and the inclusion of a passive communication condition.

The consistently low-performing profile in Study 3, Silent Overachievers, resembled the low cognitive ability and low trait confidence profile from Study 2. Notably, Silent Overachievers exhibited amplified performance under passive communication, demonstrating

that dyads initially classified as “poor performers” can thrive under favourable conditions. In contrast, Communication-Enhanced Underachievers underperformed in the passive communication condition, suggesting that even highly capable dyads may struggle when the communication mode does not align with their strengths.

Together, the latent profiles identified across Studies 2 and 3 demonstrate consistent and interpretable patterns of dyadic functioning. These profiles highlight the critical interplay between cognitive ability, trait confidence, and communication modality. They also illustrate that different dyads benefit from different forms of communication, and these benefits are shaped systematically by individual differences in trait confidence.

5.3 Key Contributions

5.3.1 Extension of the Confidence Theory

The confidence theory proposes that group members use each other's expressed decision confidence as a heuristic for inferring response accuracy. This reliance on confidence often underpins the 2HBT1 effect, where groups outperform individuals. However, prior research has largely treated individual differences in trait confidence as statistical noise, rather than a source of meaningful variability with theoretical implications. In contrast, this thesis demonstrated that trait confidence plays an important role in shaping dyadic decision processes and outcomes for static well-structured tasks. Our findings show that trait confidence influences: 1) the degree to which dyads rely on decision confidence during collaboration; 2) the extent of confidence matching between dyad members; 3) the magnitude of dyadic confidence gains in the active communication condition; and 4) the size of dyadic accuracy gains across different communication modalities.

By establishing trait confidence as a stable influence in dyadic decision-making, this thesis extends the confidence theory in several important ways. It highlights that the

informational value of expressed confidence is not purely situational but is systematically moderated by individual differences. This novel reframing strengthens the theoretical foundation of the confidence theory and opens new areas for research into when and how confidence serves as a useful cue for collective decision-making.

5.3.2 Person-Centred Statistical Analyses

This thesis employed a diverse set of statistical methods to investigate dyadic decision-making. Consistent with conventions in group decision-making research, many analyses adopted a variable-centred approach, including, ANOVA, multiple regression, LMM, and CFA.

A key contribution of this thesis was applying a person-centred approach, using LPA to identify psychological profiles of dyadic performance. The use of both variable-centred and person-centred approaches is critical for establishing the robustness and ecological validity of findings (Botvinik-Nezer et al., 2020). The convergence of results across these analytic approaches provides compelling evidence that the findings we observed are not artifacts of methodological choices but instead reflect naturally occurring processes and outcomes for dyads.

5.3.3 Use of Different Tasks and Response Options

Another strength of this thesis was the deliberate use of varied task types and response formats across studies. Study 1 employed a dynamic driving simulation that required real-time adaptation to environmental changes and engaged complex cognitive-motor coordination. Study 2 employed a battery of static, well-structured tasks with differing response formats: three tasks used multi-alternative forced-choice responses (with two, four, or eight options), and one required free-text responses. Study 3 employed two-alternative forced-choice general knowledge tests, allowing control over item difficulty.

Despite these methodological differences, the findings for dyadic accuracy were consistent: under stable, static task conditions in all three studies, dyads outperformed individuals, on average. This consistency across different task types and response formats supports the generalisability of the primary findings and demonstrates the robustness of the 2HBT1 effect across a broad range of decision-making contexts.

5.4 General Limitations

5.4.1 The Unit of Analysis Problem

When groups make decisions, a hierarchical structure emerges in which individual characteristics and processes combine to form group-level properties, and both levels influence groups outcomes. For this reason, a multilevel modelling approach is typically recommended to account for the interdependence of observations within dyads (Gonzalez & Griffin, 2023; Kenny et al., 2006). However, such models require sufficient variability in outcomes and predictors at both the within-dyad and between-dyad levels.

In Studies 1 and 2, dyadic responses were measured solely at the dyad level, meaning there was no within-dyad variance in accuracy (or speed in Study 1). As Kenny & Kashy (2014; p. 591) note, “if only a single outcome is obtained for each group, then group is treated as the unit of analysis, and special analytic methods are not required.” Accordingly, it was appropriate to treat the dyad as the unit of analysis without employing multilevel modelling. However, individual-level predictors such as cognitive ability and social sensitivity were analysed by aggregating across the two dyad members (by computing the mean). While this reduces within-dyad variability for individual measures, it is widely accepted in group decision-making research (e.g., Blanchard et al., 2020; Koriat, 2015; Schuldt et al., 2017). Nonetheless, aggregation can obscure meaningful within-dyad differences and should be applied cautiously, particularly when dyad members differ substantially on key individual characteristics.

In Study 3, “dyadic” outcomes were measured at both the individual level (in the isolated and passive communication conditions) and the dyad level (in the active communication condition). This hybrid design allowed the data to be modelled at the individual level while accounting for the nested structure of dyad membership. However, it also introduced a different trade-off: in the active communication condition, each dyad produced only one collective response per item, but this response was duplicated across both members in the dataset. This artificially inflated the sample size for dyadic responses in the active condition from N = 105 to N = 210 without adding new information, potentially reducing standard errors and increasing the risk of Type I errors.

Each analytic decision involved specific trade-offs. Aggregating individual-level predictors (as in Study 2) may have suppressed meaningful within-dyad differences, while duplicating dyad-level responses (Study 3) may have artificially reduced standard errors. To mitigate these risks in Study 3, we used LMMs with appropriate random effects structures that accounted for the nested nature of the data, helping to preserve the validity of our statistical inferences. Regardless of whether analyses are conducted at the individual or dyad level, there are inherent trade-offs. A strength of our approach is that it employed both levels of analysis and found consistent patterns across approaches. Nonetheless, future studies should systematically compare analytic strategies to determine the most appropriate way to represent individual-level constructs within dyadic contexts.

Finally, another important unit of analysis, the item level, was only partially explored. While we examined item-level dynamics in our confidence matching analyses (Study 3), the other analyses were conducted at the variable or person level, aggregating across items. Broader application of item-level analyses could yield additional insights into dynamic processes occurring within dyads and should be considered a valuable direction for future research.

5.4.2 Limited Generalisability Based on Group Size

This thesis focused exclusively on dyads, which limits the generalisability of the findings to larger groups. As group size increases, the nature of collaboration changes. Social factors, such as those identified by Woolley et al.'s (2010), including social sensitivity and equality of turn-taking, may become more influential due to the increasing complexity of coordination and interaction processes.

Although prior research has shown that confidence is influential on decisions made by larger groups (e.g., Price & Stone, 2004; Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997), it remains unclear whether the degree of reliance on confidence shifts as group size increases and other decision strategies, such as the plurality rule, emerge. Supporting this, Blanchard et al. (2024) found in a simulation study that as group size increased, confidence became a less effective decision heuristic, while using the plurality rule (i.e., relying on the most common response) became more beneficial. These findings suggest that the influence of confidence observed in dyads may decrease in larger groups. Future research should explore how both trait confidence and decision confidence influence group decision-making across a broader range of group sizes.

5.4.3 Limited Consideration of Gender Effects

This thesis did not aim to examine gender differences, despite growing evidence that gender composition can influence group processes and outcomes (e.g., Carli, 2001; Woolley et al., 2010). Some prior studies have controlled for gender by recruiting only same-gender dyads (e.g., Bahrami et al., 2010; Bang et al., 2014). Woolley at al. (2010), for instance, found that a higher proportion of female group members was associated with greater collective intelligence and group performance. In contrast, this thesis observed the opposite pattern: dyads with a higher proportion of females tended to perform worse than those with more males in study 1 (Chapter 2) and study 2 (Chapter 3). In study 2 a more nuanced pattern

emerged. While a positive indirect effect was observed (mediated by social sensitivity and consistent with Woolley et al., 2010), it was outweighed by a stronger negative direct effect indicating that the proportion of females was associated with lower performance.

These diverging findings may reflect differences in task demands, with subtle gender-based differences in relevant cognitive abilities contributing to the outcomes. For example, tasks employed in study 2 relied on fluid reasoning, a domain where males tend to show a small advantage (e.g., Halpern et al., 2007). Whereas verbal abilities may have played a greater role in Woolley et al.'s (2010) tasks which tended to be more open ended and subjective (ill-structured tasks). Females typically have a small advantage on verbal abilities (e.g., Reilly et al., 2019). Additionally, the use of university student samples, where females comprised the majority of participants in each study (ranging from 63%-75%) may further limit the generalisability of our findings. Although gender effects were beyond the scope of this thesis, future research should explore how gender interacts with trait confidence to shape dyadic decision-making.

5.4.4 Limited Generalisability Based on Task Types

While Studies 1 to 3 collectively employed a mix of dynamic and static tasks, our effects should be examined across a broader range of domains to assess their generalisability.

In Study 1, we employed a driving simulation, a dynamic task requiring real-time integration of cognitive, motor, and social processes under changing environmental conditions. While other dynamic tasks (e.g., air traffic control simulations, emergency medical response, or disaster management scenarios) share these characteristics, they may place differing demands on specific cognitive or motor abilities. Future research should examine the 2HBT1 effect and the role of communication quality in dynamic tasks that tap into different combinations of abilities.

In Study 2 and 3, we used static well-structured tasks that predominantly engaged fluid reasoning and crystallised intelligence. Future research should investigate whether the effects of trait confidence and communication also extend to tasks involving different cognitive abilities such as visual perception, processing speed, and short-term memory. Similarly, these effects should be explored for ill-structured tasks, which feature multiple correct solutions and multiple strategic pathways to completion, and misleading tasks where the confidence accuracy relationship is reversed. Exploring these domains would clarify whether the patterns identified in this thesis are specific to certain cognitive contexts or task characteristics or reflect general principles of dyadic collaboration.

5.5 Implications and Future Directions

5.5.1 Theoretical Implications

This thesis extends the confidence theory by demonstrating that trait confidence systematically influences dyadic decision-making processes and outcomes. Previous formulations of the confidence theory focused on decision confidence expressed in the moment. The findings presented here show that stable individual differences in trait confidence also shape how dyads express, use, and align their decision confidence, and achieve performance gains. This expanded view reframes models of collective metacognition, placing trait confidence alongside decision confidence as a core determinant of collective processes and outcomes. Furthermore, the results of this thesis challenge interpretations of collective intelligence that focus primarily on social factors, such as social sensitivity and communication equality (e.g., Woolley et al., 2010). While social processes are important, our findings show that cognitive ability and strategic, context sensitive communication are more critical ingredients for effective collaboration, especially in dyads. It is likely that these factors remain influential in larger groups. Overall, theoretical frameworks seeking to explain

dyadic decision-making should integrate cognitive, metacognitive, and social factors to provide a more comprehensive account of how they function.

5.5.2 Methodological Implications

This thesis also employed novel methodological approaches to the study of dyadic decision-making. By combining variable-centred and person-centred analyses, it provides a nuanced understanding of how individual and dyadic characteristics interact to influence outcomes. The application of LPA allowed for the identification of psychological profiles that would have been obscured in variable-centred analyses alone. Additionally, modelling dyadic and individual level data, while navigating inherent trade-offs illustrated the importance of considering multiple levels of analysis when studying collaboration. Future studies should continue to employ multi-level modelling strategies and broaden the use of item-level analyses to capture the dynamic, evolving processes within groups. Item-level analyses could further illuminate how decision-making processes shift as a function of trait confidence across trials, tasks, or changing environmental conditions. Additionally, systematic comparisons between individual-level and dyad-level data structures remain an important area for refining best practices in group research.

5.5.3 Selection and Training

The findings of this thesis also offer practical insights for forming non-hierarchical (leaderless) dyads that are fit for context. For dynamic tasks, the use of dyads should be informed by the performance goals and environmental conditions. Under stable operational conditions, dyads show an accuracy advantage, whereas under volatile operational conditions that rapidly change, dyads lose their advantage and become slower than individuals. For example, during a live patrol of unfamiliar terrain, a driver and navigator operating as peers may work more effectively under stable conditions but when rapid route changes are needed due to unexpected obstacles or threats, the time required to coordinate decisions can hinder

performance compared to a driver acting alone. Thus, the use of dyads in dynamic tasks should be carefully considered; individuals may be better suited if the environmental conditions are volatile.

For static, well-structured tasks, trait confidence emerged as a key factor that can be targeted when pairing individuals but again, context matters. Passive communication conditions may be preferred for low-trait confidence dyads due to lower time and resource costs without diminishing accuracy gains. Whereas active communication conditions are more beneficial for mixed-trait and high-trait confidence dyads, as verbal discussion enables more effective identification of the most accurate judgment. For example, during a map-based intelligence task requiring two analysts to identify potential threats, pairing a highly confident individual with a less confident peer may result in greater accuracy gains when they communicate verbally, however, two low confidence individuals may achieve an equivalent benefit from simple non-verbal and non-interactive communication. Selection strategies should be sensitive to both trait confidence and the communication context.

Although this thesis did not examine the effects of training, its findings support several suggestions. Future research should examine whether training dyads to be aware of confidence biases (e.g., stable overconfidence) can enhance collective performance. This may reduce the influence of confidence by altering the weights individuals place on confidence judgments, but it may also produce unintended effects on dyadic processes and outcomes. Thus, research should explore its efficacy before recommending training protocols.

Finally, communication training should move beyond promoting equality of participation and instead encourage strategic communication tailored to the dyad's composition, task demands, and assigned roles. For example, in Study 1 (Chapter 2), where dyad members had distinct roles (i.e., driver and navigator), only the quality of the navigator's communication was associated with task accuracy (measured by collisions),

while the driver's communication quality was not. Teaching individuals to adapt their communication style based on the context, their partner's role, and their partner's characteristics (as observed in Study 2 and Study 3) could enhance dyadic outcomes.

5.6 Conclusion

This thesis makes substantial contributions to understanding dyadic groups decision-making by systematically examining how trait confidence and communication shape collaborative processes and outcomes. Across three empirical studies, these relationships were investigated under varying task and communication conditions. Several novel contributions to the group decision-making literature emerged. First, this thesis developed a method for quantifying the communication quality during dynamic tasks. Second, it introduced a methodological and analytical framework for evaluating dyadic collective intelligence. Third, it was the first to systematically investigate how trait confidence influences dyadic outcomes under different modes of communication. Finally, it provided the first empirical test of whether confidence matching arises spontaneously in consensus-seeking dyads engaged in verbal discussion.

Together, the studies identified distinct profiles of dyadic performance and demonstrated that no single factor (i.e., cognitive ability, trait confidence, or communication) acts in isolation. Instead, collaborative success depends on a dynamic interplay between these factors, with important implications for theory and practice. Based on the evidence from this thesis, the key recommendation for improving dyadic decision-making in applied contexts is to pair individuals based on their individual traits and task demands. For example, dyads composed of high-trait confidence individuals may benefit most from verbal communication, while lower-trait confidence dyads may benefit under more structured non-verbal communication conditions that remove the demands of real-time discussion. Additionally, communication training should go beyond promoting equality and instead emphasise

adaptive communication strategies that are sensitive to role asymmetries and partner characteristics.

Overall, this thesis highlights the critical roles of trait confidence and communication in shaping dyadic decision-making processes and outcomes. It opens new directions for future research to explore how stable individual differences influence collaboration.

References

Abbott, E. F., Laack, T. A., Licatino, L. K., Wood-Wentz, C. M., Warner, P. A., Torsher, L. C., ... & Rieck, K. M. (2021). Comparison of dyad versus individual simulation-based training on stress, anxiety, cognitive load, and performance: a randomized controlled trial. *BMC medical education*, 21, 1-10. https://doi.org/10.1186/s12909-021-02786-6

Aggarwal, I., Woolley, A. W., Chabris, C. F., & Malone, T. W. (2019). The Impact of Cognitive Style Diversity on Implicit Learning in Teams. *Frontiers in Psychology*, 10(112). https://doi.org/10.3389/fpsyg.2019.00112

Albolino, S., Cook, R., & O'Connor, M. (2007). Sensemaking, safety, and cooperative work in the intensive care unit. *Cognition, Technology & Work*, 9, 131–137. https://doi.org/10.1007/s10111-006-0057-5

Algesheimer, R., Dholakia, U. M., & Gurău, C. (2011). Virtual team performance in a highly competitive environment. *Group & Organization Management*, 36, 161–190. https://doi.org/10.1177/1059601110391251

Anderson, K. J., & Leaper, C. (1998). Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*, 39(3-4), 225-252. https://doi.org/10.1023/A:1018802521676

Anstey, K. J., Wood, J., Lord, S. & Walker, J. G. (2005). Cognitive, sensory and physical factors enabling driving safety in older adults. *Clinical Psychology Review*, 25, 45–65. https://doi.org/10.1016/j.cpr.2004.07.008

Asimakopulos, J., Boychuck, Z., Sondergaard, D., Poulin, V., Ménard, I., & Korner-Bitensky, N. (2012). Assessing executive function in relation to fitness to drive: A review of tools and their ability to predict safe driving. *Australian Occupational Therapy Journal*, *59*(6), 402-427. https://doi.org/10.1111/j.1440-1630.2011.00963.x

Askay, D., Metcalf, L., Rosenberg, L., & Willcox, G. (2019). *Enhancing Group Social Perceptiveness through a Swarm-based Decision-Making Platform.* Paper presented at the Proceedings of the 52nd Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2019.061

Baddeley, A. (1992). Working Memory. *Science.* 255, 556–559. https://doi.org/10.1126/science.1736359

Bahrami, B., Didino, D., Frith, C., Butterworth, B., & Rees, G. (2013). Collective enumeration. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 338–347. https://doi.org/10.1037/a0029717

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012a). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1350-1365. https://doi.org/10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G. & Frith, C. (2012b). Together, Slowly but Surely. *Journal of Experimental Psychology: Human Perception and Performance*, 38 (1), 3-8. https://doi.org/10.1037/a0025708.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*(5995), 1081-1085. https://doi.org/10.1126/science.1185718

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., ... & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 0117. https://doi.org/10.1038/s41562-017-0117

Bang, D., Fusaroli, R., Tylen, K., Olsen, K., Latham, P., Lau, J., ... Bahrami, B. (2014). Does interaction matter? Testing whether a Confidence heuristic can replace interaction in

collective decision-making. *Consciousness and Cognition*, *26*, 13–23. https://doi.org/10.1016/j.concog.2014.02.002

Barlow, J. B., & Dennis, A. R. (2016). Not as smart as we think: A study of collective intelligence in virtual groups. *Journal of Management Information Systems*, 33(3), 684-712. https://doi.org/10.1080/07421222.2016.1243944

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175. https://doi.org/10.1023/b:jadd.0000022607.19833.00

Barrick, M. R., Stewart, G. L., Neubert, M. J., and Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, 83, 377. https://doi.org/10.1037/0021-9010.83.3.377

Bates, T. C., & Gupta, S. (2017). Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups. *Intelligence*, *60*, 46-56. https://doi.org/10.1016/j.intell.2016.11.004

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Beaubien, J., & Baker, D. (2004). The use of simulation for training teamwork skills in health care: how low can you go? *Quality and Safety in Health Care*, *13*(suppl 1), i51–6. https://doi.org/10.1136/qshc.2004.009845

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: a meta-analysis. *Journal of Applied Psychology*, *92*(3), 595. https://doi.org/10.1037/0021-9010.92.3.595

Bell, B. S., & Kozlowski, S. W. (2002). A typology of virtual teams: Implications for effective leadership. *Group & Organization Management.* 27, 14–49. https://doi.org/10.1177/1059601102027001003

Black, J. E. (2018). An IRT Analysis of the Reading the Mind in the Eyes Test. *Journal of Personality Assessment*, *101*(4), 425–433. https://doi.org/10.1080/00223891.2018.1447946

Blanchard, M. D., Jackson, S. A., & Kleitman, S. (2020). Collective decision making reduces metacognitive control and increases error rates, particularly for overconfident individuals. *Journal of Behavioral Decision Making*, *33*(3), 348-375. https://doi.org/10.1002/bdm.2156

Blanchard, M. D., Herzog, S. M., Kämmer, J. E., Zöller, N., Kostopoulou, O., & Kurvers, R. H. (2024). Collective Intelligence Increases Diagnostic Accuracy in a General Practice Setting. *Medical Decision Making*, Advance online publication. https://doi.org/10.1177/0272989X241241001

Blanchard, M. D., Kleitman, S., & Aidman, E. (2023). Are two naïve and distributed heads better than one? Factors influencing the performance of teams in a challenging real-time task. *Frontiers in Psychology*, *14*, 1042710. https://doi.org/10.3389/fpsyg.2023.1042710

Bosch-Sijtsema, P. M., Fruchter, R., Vartiainen, M., & Ruohomäki, V. (2011). A framework to analyze knowledge work in distributed teams. *Group & Organization Management*, *36*(3), 275-307. https://doi.org/10.1177/1059601111403625

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature,* 582, 84–88 (2020). https://doi.org/10.1038/s41586-020-2314-9

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human factors*, *40*(4), 672-679. https://doi.org/10.1518/001872098779649265

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, *67*(1), 59-77. https://doi.org/10.1016/j.jml.2012.04.002

Broadway, J. M., and Engle, R. W. (2010). Validating running memory span: measurement of Working Memory capacity and links with fluid intelligence. *Behavioral Research Methods 42*, 563–570. https://doi.org/10.3758/BRM.42.2.563

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92*(5), 938. https://doi.org/10.1037/0022-3514.92.5.938

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2012). Explaining adult age differences in decision-making competence. *Journal of Behavioral Decision Making*, *25*(4), 352-360. https://doi.org/10.1002/bdm.712

Campbell, W. K., Bonacci, A. M., Shelton, J., Exline, J. J., & Bushman, B. J. (2004). Psychological Entitlement: Interpersonal Consequences and Validation of a Self-Report Measure. *Journal of Personality Assessment*, *83*(1), 29–45. https://doi.org/10.1207/s15327752jpa8301_04

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan Jr. (Ed.), *Individual and group decision making: Current issues* (pp. 221–246). Hillsdale, NJ: Erlbaum.

Carli, L. L. (2001), 'Assertiveness', in J. Worell (ed.), *Encyclopedia of Women and Gender: Sex Similarities and Differences and the Impact of Society on Gender*. San Diego, CA: Academic Press.

Carli, L. L. and Bukatko, D. (2000), 'Gender, communication, and social influence: A developmental perspective', in T. Eckes and H. M. Trautner (eds), *The Developmental Social Psychology of Gender*. Mahwah, NJ: Lawrence Erlbaum Associates. Pp. 295–331.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge, England: Cambridge University Press.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, 67(2), 319–333. https://doi.org/10.1037/0022-3514.67.2.319

Choi, H. S., & Levine, J. M. (2004). Minority influence in work teams: The impact of newcomers. *Journal of experimental social psychology*, *40*(2), 273-280. https://doi.org/10.1016/S0022-1031(03)00101-X

Clark, S. L., & Muthén, B. O. (2009). *Relating latent class analysis results to variables not included in the analysis*. Unpublished manuscript.

Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management*, *23*(3), 239-290. https://doi.org/10.1177/014920639702300303

Cohen, S., & Cohen, S. (1980). Aftereffects of stress on human performance and social behavior: a review of research and theory. *Psychological Bulletin*, *88*(1), 82–108. https://doi.org/10.1037/0033-2909.88.1.82

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depression and anxiety*, *18*(2), 76-82. https://doi.org/10.1002/da.10113

Cooke, N. J. (2015). Team cognition as interaction. *Current directions in psychological science*, *24*(6), 415-419. https://doi.org/10.1177/0963721415602474

Cooke, N., Gorman, J., Duran, J., Taylor, A., & Cooke, N. (2007). Team cognition in experienced command-and-control teams. *Journal of Experimental Psychology. Applied*, *13*(3), 146–157. https://doi.org/10.1037/1076-898X.13.3.146

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, 37, 255–285. https://doi.org/10.1111/cogs.12009

Cooke, N. J., Gorman, J. C., & Rowe, L. J. 2009. An ecological perspective on team cognition. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations* (pp. 157-182). New York: Routledge, Taylor & Francis Group.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in Measuring Team Cognition. In E. Salas & S. M. Fiore (Eds.), *Team Cognition: Understanding the Factors that Drive Process and Performance* (pp. 83-106). Washington, D.C.: American Psychological Association.

Credé, M., Howardson, G. (2017). The structure of group task performance – A second look at "collective intelligence": Comment on Woolley et al. (2010). *Journal of Applied Psychology*, *102*, 1483–1492. https://doi.org/10.1037/apl0000176

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. https://doi.org/10.1007/BF02310555

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25*, 331-351. https://doi.org/10.1002/bdm.731

Devine, D. J. (2002). A review and integration of classification systems relevant to teams in organizations. *Group Dynamics*, *6*, 291–310. https://doi.org/10.1037/1089-2699.6.4.291

Devine, D. J., and Philips, J. L. (2001). Do smarter teams do better: A meta- analysis of cognitive ability and team performance. *Small Group Research, 32*, 507–532. https://doi.org/10.1177/104649640103200501

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment*, *18*(2), 192. https://doi.org/10.1037/1040-3590.18.2.192

Driskell, J. E., and J. H. Johnston. (1998). "Stress Exposure Training." In *Making Decisions under Stress – Implications for Individual and Team Training*, edited by J. A. Cannon-Bowers and E. Salas, pp. 191–217. Washington, DC: American Psychological Association.

Driskell, J., Salas, E., & Johnston, J. (1999). Does stress lead to a loss of team perspective? *Group Dynamics: Theory, Research, and Practice*, *3*(4), 291–302. https://doi.org/10.1037/1089-2699.3.4.291

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative science quarterly*, *44*(2), 350-383. https://doi.org/10.2307/2666999

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, 20, 17–28. https://doi.org/10.1016/S0268-4012(99)00051-1

Ellis, A. (2006). System breakdown: the role of mental models and transactive memory in the relationship between acute stress and team performance. *Academy of Management Journal*, *49*, 576–589. https://doi.org/10.5465/AMJ.2006.21794674

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting,* 32, 97-101. Sage CA: Los Angeles, CA: Sage Publications.

Endsley, M. R. (2021). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*, 63, 124–150. https://doi.org/10.1177/0018720819875376

Engel, D., Woolley, A. W., Aggarwal, I., Chabris, C. F., Takahashi, M., Nemoto, K., ... & Malone, T. W. (2015, April). Collective intelligence in computer-mediated collaboration emerges in different contexts and cultures. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3769-3778). https://doi.org/10.1145/2702123.2702259

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), 1–16. https://doi.org/10.1371/journal.pone.0115212

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, *16*(1), 143-149. https://doi.org/10.3758/BF03203267

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, 7(2), 117-140. https://doi.org/10.1177/001872675400700202

Fischer, U., McDonnell, L., & Orasanu, J. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation, space, and environmental medicine*, *78*(5), B86-B95.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme Confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552–564. https://doi.org/10.1037/0096-1523.3.4.552

Fishman, P. M. (1978), 'Interaction: The work women do', *Social Problems, 25*, 397–406. https://doi.org/10.2307/800492

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–236). Hillsdale, NJ: Erlbaum.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378. https://doi.org/10.1037/h0031619

Foda, H., Barger, K., Navajas, J., & Bahrami, B. (2017). Domain-general idiosyncratic anchoring of metacognition. Unpublished manuscript.

Foushee, H. C. (1984). Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. *American Psychologist*, 39, 885–893. https://doi.org/10.1037/0003-066X.39.8.885

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42. https://doi.org/10.1257/089533005775196732

Friedemann, M., Bang, D., & Yeung, N. (2024). Normative and informational confidence matching. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001706

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

Gardner, A. K., Scott, D. J., & AbdelFattah, K. R. (2017). Do great teams think alike? An examination of team mental models and their impact on team performance. *Surgery*, 161, 1203-1208. https://doi.org/10.1016/j.surg.2016.11.010

Geil, D. M. M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, 4, 231-248. https://doi.org/10.1080/135467898394148

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gladstein, D., & Reilly, N. (1985). Group decision making under threat: The tycoon game. *Academy of Management Journal*, *28*, 613–627. https://doi.org/10.2307/256117

Glynn, S. J., & Henning, R. A. (2000). Can teams outperform individuals in a simulated dynamic control task?. In *Proceedings of the human factors and ergonomics society annual meeting 44*, 6-141. Sage CA: Los Angeles, CA: SAGE Publications.

Gonzalez, R., & Griffin, D. (2023). Dyadic data analysis. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Data analysis and research publication* (2nd ed., pp. 465–478). American Psychological Association. https://doi.org/10.1037/0000320-021

González-Romá, V., & Hernández, A. (2014). Climate uniformity: Its influence on team communication quality, task conflict, and team performance. *Journal of Applied Psychology*, 99, 1042–1058. https://psycnet.apa.org/doi/10.1037/a0037868

Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, *7*(5), 398. https://doi.org/10.1037/h0074666

Gorman, J. C., Cooke, N. J., Amazeen, P. G., & Fouse, S. (2012). Measuring patterns in team interaction sequences using a discrete recurrence approach. *Human Factors*, *54*(4), 503-517. https://doi.org/10.1177/0018720811426140

Gorman, J. C., Cooke, N. J., Pederson, H. K., & DeJoode, J. A. (2005). Coordinated awareness of situation by teams (CAST): Measuring team situation awareness of a communication glitch. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49, 274-277. Sage CA: Los Angeles, CA: SAGE Publications.

Graefe, A., & Armstrong, J. S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting*, *27*, 183-195. https://doi.org/10.1016/j.ijforecast.2010.05.004

Graf, V., Gimpel, H., & Barlow, J. B. (2019). Clarifying the Structure of Collective Intelligence in Teams: A Meta-Analysis. In *Proceedings of Collective Intelligence Conference*.

Graf-Drasch, V., Gimpel, H., Barlow, J. B., & Dennis, A. R. (2022). Task structure as a boundary condition for collective intelligence. *Personnel Psychology*, *75*(3), 739-761. https://doi.org/10.1111/peps.12489

Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill Book Co., Inc.

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. *Advances in Experimental Social Psychology*, *8*, 45-99. https://doi.org/10.1016/S0065-2601(08)60248-8

Hackman, J. R., & Vidmar, N. (1970). Effects of size and task type on group performance and member reactions. *Sociometry*, 37-54. https://doi.org/10.2307/2786271

Haesevoets, T., Reinders Folmer, C., Bostyn, D. H., & Van Hiel, A. (2018). Behavioural Consistency within the Prisoner's Dilemma Game: The Role of Personality and Situation. *European Journal of Personality*, 32(4), 405-426. https://doi.org/10.1002/per.2158

Haesevoets, T., Van Hiel, A., Van Assche, J., Bostyn, D. H., & Folmer, C. R. (2019). An exploration of the motivational basis of take-some and give-some games. *Judgment and Decision Making*, 14(5), 534-546. https://doi.org/10.1017/S1930297500004836

Haigney, D. E., Taylor, R. G., & Westerman, S. J. (2000). Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes. *Transportation Research Part F: Traffic Psychology and Behaviour*, 3, 113-121. https://doi.org/10.1016/S1369-8478(00)00020-6

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1-51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2009). Mental state decoding in past major depression: Effect of sad versus happy mood induction. *Cognition and Emotion*, *24*(3), 497–513. https://doi.org/10.1080/02699930902750249

Hart, S. G., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier.

Hastie, R., & Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. *Psychological Review*, 112(2), 494–508. https://doi.org/10.1037/0033-295X.112.2.494

Helmreich, R., Merritt, A., & Wilhelm, J. (1999). The Evolution of Crew Resource Management Training in Commercial Aviation. *The International Journal of Aviation Psychology*, *9*(1), 19–32. https://doi.org/10.1207/s15327108ijap0901_2

Henry, R. A. (1993). Group judgment accuracy: Reliability and validity of postdiscussion Confidence judgments. *Organizational Behavior and Human Decision Processes*, *56*, 11–27. https://doi.org/10.1006/obhd.1993.1043

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modelling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 202-226. https://doi.org/10.1080/10705510709336744

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin*, *91*(3), 517–539. https://doi.org/10.1037/0033-2909.91.3.517

Hill, S.G., Byers, J.C., Zaklad, A.L., Christ, R.E., & Bittner, A.C. (1988). Workload assessment of a mobile air defences system. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 1068–1072). Santa Monica, CA: Human Factors Society.

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology*, *59*(4), 705–718. https://doi.org/10.1037/0022-3514.59. 4.705

Hinsz, V.B., Tindale, R.S., & Vollrath, D.A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, 121, 43-64. https://doi.org/10.1037/0033-2909.121.1.43

Hirst, G., & Mann, L. (2004). A model of R & D leadership and team communication: The relationship with project performance. *R & D Management*, 34, 147–160. https://doi.org/10.1111/j.1467-9310.2004.00330.x

Hollingshead, A., & Poole, M. S. (2012). *Research Methods for Studying Groups: A Guide to Approaches, Tools, and Technologies*. Hoboken: Taylor and Francis, 2012.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, 92, 1644-1655. https://doi.org/10.1257/000282802762024700

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5). https://doi.org/10.1037/h0023816

Ifenthaler, D., Eseryel, D., & Ge, X. (2012). Assessment in game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 1-8). New York: Springer.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, 56: 517-543. https://doi.org/10.1146/annurev.psych.56.091103.070250

Imbimbo, E., Stefanelli, F., & Guazzini, A. (2020). Adolescent's collective intel- ligence: Empirical evidence in real and online classmates groups. *Future Internet, 12*(5), 81. https://doi.org/10.3390/fi12050081

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*(4), 505-524. https://doi.org/10.1348/000712605X53542

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and Confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning, 9*(1), 25–49. https://doi.org/10.1007/s11409-013-9110-y

Jackson, S. A., Kleitman, S., & Aidman, E. (2014). Low cognitive load and reduced arousal impede practice effects on executive functioning, metacognitive confidence and decision making. *PloS one*, *9*(12), e115689. https://doi.org/10.1371/journal.pone.0115689

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2016a). Decision pattern analysis as a general framework for studying individual differences in decision making. *Journal of Behavioral Decision Making*, *29*(4), 392-408. https://doi.org/10.1002/bdm.1889

Jackson, S. A., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence, and control thresholds explain individual differences in heuristics and biases. *Frontiers in Psychology*, *7*, 1559. https://doi.org/10.3389/fpsyg.2016.01559

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2017). Individual differences in decision making depend on cognitive abilities, monitoring and control. *Journal of Behavioral Decision Making*, *30*(2), 209-223. https://doi.org/10.1002/bdm.1939

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. Free press.

Johnson, D. M. (1939). Confidence and speed in the two-category judgement. Archives of Psychology, 241,1-52.

Kameda, T., Toyokawa, W. & Tindale, R.S. Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology 1*, 345–357 (2022). https://doi.org/10.1038/s44159-022-00054-y

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of Working Memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189. https://doi.org/10.1037/0096-3445.133.2.189

Kanki, B. G., Helmreich, R. L., & Anca, J. (Eds.). (2010). *Crew resource management*. Academic Press. Kirkman.

Kenny, D. A., & Kashy, D. A. (2014). The Design and Analysis of Data from Dyads and Groups. In Reis, H. T., & Judd, C. M. (Eds.), *Handbook of Research Methods in*

Social and Personality Psychology (pp. 589–607). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.027

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Guilford Press.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623-655. https://doi.org/10.1146/annurev.psych.55.090902.142009

Keyton, J., Beck, S. J., & Asbury, M. B. (2010). Macrocognition: a communication perspective. *Theoretical Issues in Ergonomics Science*, 11, 272-286. https://doi.org/10.1080/14639221003729136

Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What Makes a Strong Team?: Using Collective Intelligence to Predict Team Performance in League of Legends. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2316–2329. https://doi.org/10.1145/2998181.2998185

Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the Mind's Eye: A Meta-Analytic Investigation of the Nomological Network and Internal Consistency of the "Reading the Mind in the Eyes" Test. *Assessment*, *29*(5), 872-895. https://doi.org/10.1177/1073191121996469

Kline, P. (2014). *An Easy Guide to Factor Analysis*. Routledge. https://doi.org/10.4324/9781315788135

Kleitman, S., Jackson, S. A., Zhang, L. M., Blanchard, M., Rizvandi, N. B. & Aidman, E. (2022). Applying evidence-centered design for the measurement of psychological

resilience: The development and preliminary validation of a novel simulation-based assessment methodology. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.717568

Kleitman, S., Zhang, L. M., Blanchard, M. D., Law, M. K. H., Xiong, Z., Jackson, S. A., & Aidman, E. (2020). A 'Maze Runner' design for the measurement of resilience and adaptability: Validation of a novel simulation-based assessment methodology. Unpublished manuscript.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*(3), 321-341. https://doi.org/10.1002/acp.705

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*(2), 161-173. https://doi.org/10.1016/j.lindif.2007.03.004

Koriat, A. (2008). Subjective Confidence in one's answers: The con- sensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2011). Subjective Confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140*, 117–139. https://doi.org/10.1037/a0022171

Koriat, A. (2012a). When are two heads better than one and why? *Science*, *336*, 360–362. https://doi.org/10.1126/science.1216549

Koriat, A. (2012b). The self-consistency model of subjective Confidence. *Psychological Review, 119*, 80–113. https://doi.org/10.1037/a0025648

Koriat, A. (2012c). The relationships between monitoring, regulation and performance. *Learning and Instruction, 22*(4), 296-298. https://doi.org/10.1016/j.learninstruc.2012.01.002

Koriat, A. (2015). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*, 934–950. https://doi.org/10.1037/xge0000092

Koriat, A. (2024). Subjective Confidence as a Monitor of the Replicability of the Response. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231224387

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517. https://doi.org/10.1037/0033-295X.103.3.490

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for Confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Kozlowski, S. W. J., & Bell, B. S. (2013). Work groups and teams in organizations. In N. W. Schmitt, S. Highhouse, & I. B. Weiner (Eds.), *Handbook of psychology: Industrial and organizational psychology* (pp. 412–469). John Wiley & Sons, Inc..

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3- 90). San Francisco, CA: Jossey-Bass.

Krekels, G., & Pandelaere, M. (2015). Dispositional greed. *Personality and Individual Differences*, 74, 225–230. https://doi.org/10.1016/j.paid.2014.10.036

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and quantity*, *38*, 787-800. https://doi.org/10.1007/s11135-004-8107-7

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., ... & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777-8782. https://doi.org/10.1073/pnas.1601827113

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Laapotti, S., & Keskinen, E. (2004). Has the difference in accident patterns between male and female drivers changed between 1984 and 2000?. *Accident Analysis & Prevention*, 36, 577-584. https://doi.org/10.1016/S0001-4575(03)00064-2

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, 13, 163-167. https://doi.org/10.1023/A:1023260610025

Law, M. K., Stankov, L., & Kleitman, S. (2022). I choose to opt-out of answering: Individual differences in giving up behaviour on cognitive tests. *Journal of Intelligence*, *10*(4), 86. https://doi.org/10.3390/jintelligence10040086

Leet-Pellegrini, H. M. (1980), 'Conversational dominance as a function of gender and expertise', in H. Giles, W. P. Robinson and P. M. Smith (eds), *Language: Social Psychological Perspectives*. Oxford: Pergamon. Pp. 97–104.

LePine, J. A. (2003). Team adaptation and postchange performance: effects of team composition in terms of members' cognitive ability and personality. *Journal of Applied Psychology*, *88*(1), 27-39. https://doi.org/10.1037/0021-9010.88.1.27

LePine, J. A. (2005). Adaptation of teams in response to unforeseen change: effects of goal difficulty and team composition in terms of cognitive ability and goal orientation. *Journal of Applied Psychology*, *90*(6), 1153. https://doi.org/10.1037/0021-9010.90.6.1153

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, *61*(2), 273-307. https://doi.org/10.1111/j.1744-6570.2008.00114.x

LePine, J. A., Hollenbeck, J. R., Ilgen, D. R., & Hedlund, J. (1997). Effects of individual differences on the performance of hierarchical decision-making teams: Much more than g. *Journal of Applied Psychology*, *82*(5), 803. https://doi.org/10.1037/0021-9010.82.5.803

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?. *Organizational behavior and human performance*, *20*(2), 159-183. https://doi.org/10.1016/0030-5073(77)90001-0

Lingard, L., Espin, S., Whyte, S., Regehr, G., Baker, G. R., Reznick, R., ... Grober, E. (2004). Communication failures in the operating room: An observational classification of recurrent types and effects. *Quality and Safety in Health Care*, 13, 330–334. http://dx.doi.org/10.1136/qshc.2003.008425

Littlepage, G. E. (1991). Effects of group size and task characteristics on group performance: A test of Steiner's model. *Personality and Social Psychology Bulletin*, *17*(4), 449-456. https://doi.org/10.1177/0146167291174014

Lopes, P. N., Salovey, P., & Straus, R. (2003). Emotional intelligence, personality, and the perceived quality of social relationships. *Personality and Individual Differences*, *35*(3), 641-658. https://doi.org/10.1016/S0191-8869(02)00242-8

Maciejovsky, B., Sutter, M., Budescu, D. V., & Bernau, P. (2013). Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Management Science*, 59, 1255-1270. https://doi.org/10.1287/mnsc.1120.1668

MacMillan, J., Entin, E. E., & Serfaty, D. (2004). Communication overhead: The hidden cost of team cognition. In E. Salas & S. M. Fiore (Eds.), Team cognition: Understanding the factors that drive process and performance (p. 61–82). American Psychological Association. https://doi.org/10.1037/10690-004

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., ... & Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835-3840. https://doi.org/10.1073/pnas.1421692112

Mandrik, C. A., & Bao, Y. (2005). Exploring the concept and measurement of General Risk Aversion. *Advances in Consumer Research*, 32, 531–539.

Mannix, E., & Neale, M. A. (2005). What differences make a difference? The promise and reality of diverse teams in organizations. *Psychological science in the public interest*, *6*(2), 31-55. https://doi.org/10.1111/j.1529-1006.2005.00022.x

Marks, M. A., Sabella, M. J., Burke, C. S., & Zaccaro, S. J. 2002. The impact of cross training on team effectiveness. *Journal of Applied Psychology*, 87: 3-1. https://psycnet.apa.org/doi/10.1037/0021-9010.87.1.3

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 191-225. https://doi.org/10.1080/10705510902751010

Mathias, J. L. & Lucas, L. K. (2009). Cognitive predictors of unsafe driving in older drivers: A meta-analysis. *International Psychogeriatrics*, 21, 637–653. http://dx.doi.org/10.1017/S1041610209009119

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of applied psychology*, *85*(2), 273. https://doi.org/10.1037/0021-9010.85.2.273

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, 410-476. https://doi.org/10.1177%2F0149206308316061

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

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

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1-10. https://doi.org/10.1016/j.intell.2008.08.004

Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning*, *21*(3), 341–355. https://doi.org/10.1080/13546783.2014.981582

Mesmer-Magnus, J. R., DeChurch, L. A., Jimenez-Rodriguez, M., Wildman, J., & Shuffler, M. (2011). A meta-analytic investigation of virtuality and information sharing in teams. *Organizational Behavior and Human Decision Processes*, 115, 214-225. https://doi.org/10.1016/j.obhdp.2011.03.002

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology, 94*(2), 535–546. https://doi.org/10.1037/a0013773

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, *178*, 107-114. https://doi.org/10.7205/MILMED-D-13-00213

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49-100. https://doi.org/10.1006/cogp.1999.0734

Monsell, S. (2003). Task switching. *Trends in cognitive sciences*, *7*(3), 134-140. https://doi.org/10.1016/S1364-6613(03)00028-7

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502. https://psycnet.apa.org/doi/10.1037/0033-295X.115.2.502

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 26, pp. 125-173). Academic Press.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive Monitoring Accuracy and Student Performance in the Postsecondary Classroom. *The Journal of Experimental Education*, *74*(1), 7–28. http://www.jstor.org/stable/20157410

Newell, A., & Simon, H. A. (1972). Human problem solving. Prentice-Hall.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1503. https://doi.org/10.3389/fpsyg.2015.01503

Orasanu, J. (1990, July). Shared mental models and crew decision making. Paper presented at the *12th Annual Conference of the Cognitive Science Society*, Cambridge, MA.

Orasanu, J., & Salas, E. (1993). Team decision making in complex environments. In G. Klein, Orasanu, J., Calderwood, R., & Zsambok, C. (Eds.), *Decision-making in action: Models and methods* (pp. 327–345). Norwood, NJ: Ablex.

Otero, I., Martínez, A., Cuadrado, D., Lado, M., Moscoso, S., & Salgado J.F. (2024). Sex Differences in Cognitive Reflection: A Meta-Analysis. *Journal of Intelligence*, *12*(4), 39. https://doi.org/10.3390/jintelligence12040039

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614. https://doi.org/10.1016/j.intell.2022.101614

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129*(3), 257-299. https://doi.org/10.1080/00221300209602099

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review*, *21*, 1255–1262. https://doi.org/10.3758/s13423-014-0597-2

Patalano, A. L., & LeClair, Z. (2011). The influence of group decision making on indecisiveness-related decisional Confidence. *Judgment and Decision Making*, *6*(2), 163–175. https://doi.org/10.1037/e722352011-122

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, 1(4), 817. https://doi.org/10.1037/1076-8971.1.4.817

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General, 145*(8), 949–965. https://doi.org/10.1037/xge0000180

Pescetelli, N., & Yeung, N. (2020). The effects of recursive communication dynamics on belief updating. *Proceedings of the Royal Society B*, *287*(1931), 20200025. https://doi.org/10.1098/rspb.2020.0025

Pescetelli, N., & Yeung, N. (2022). Benefits of spontaneous confidence alignment between dyad members. *Collective Intelligence*, *1*(2). https://doi.org/10.1177/26339137221126915

Peeters, M. A., Van Tuijl, H. F., Rutte, C. G., and Reymen, I. M. (2006). Personality and team performance: a meta-analysis. *European Journal of Personality, 20*, 377–396. https://doi.org/10.1002/per.588

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., ... & Verbeken, B. (2025). Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self regulation* (pp. 452-502). New York: Academic Press.

Pollack, I., Johnson, L. B., and Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology, 57*, 137. https://doi.org/10.1037/h0046137

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39-57. https://doi.org/10.1002/bdm.460

Predmore, S. (1991). Microcoding of communications in accident investigation: Crew coordination in United 811 and United 232. *Proceedings of the Sixth International Symposium on Aviation Psychology*. 350-35. Columbus, OH: Ohio State University.

Prewett, M. S., Walvoord, A. A., Stilson, F. R., Rossi, M. E., & Brannick, M. T. (2009). The team personality–team performance relationship revisited: The impact of criterion

choice, pattern of workflow, and method of aggregation. *Human Performance*, *22*(4), 273-296. https://doi.org/10.1080/08959280903120253

Räder, S. B., Henriksen, A. H., Butrymovich, V., Sander, M., Jørgensen, E., Lönn, L., & Ringsted, C. V. (2014). A study of the effect of dyad practice versus that of individual practice on simulation-based complex skills learning and of students' perceptions of how and why dyad practice contributes to learning. Academic Medicine, 89(9), 1287-1294. https://10.1097/ACM.0000000000000373

Ragsdale, G., and Foley, R. A. (2011). A maternal influence on Reading the Mind in the Eyes mediated by executive function: differential parental influences on full and half-siblings. *PloS ONE*. https://doi.org/10.1371/journal.pone.0023236

Raven, J. C. (1938–65). *Progressive Matrices*. New York, NY: The Psychological Corporation.

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74*(4), 445. https://doi.org/10.1037/amp0000356

Riedl, C., Kim, Y. J., Gupta, P., Malone, T. W., & Woolley, A. W. (2021). Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, *118*(21), e2005737118. https://doi.org/10.1073/pnas.2005737118

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., & Schmidt, J. A. (2018). tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. *Journal of Open Source Software, 3*(30), 978. https://doi.org/10.21105/joss.00978

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, *100*(3), 349. https://doi.org/10.1037/0033-2909.100.3.349

Rowe, L. I., Hattie, J., & Hester, R. (2021). G versus c: comparing individual and collective intelligence across two meta-analyses. *Cognitive Research: Principles and Implications*, *6*, 1-24. https://doi.org/10.1186/s41235-021-00317-x

Rowe, L. I., Hattie, J., & Munro, J. (2024). High-performing teams: Is collective intelligence the answer?. *PLOS ONE*, *19*(8), Advance online publication. https://doi.org/10.1371/journal.pone.0307945

Salas, E., Rosen, M. A., Burke, C. S., Nicholson, D., & Howse, W. R. (2007). Markers for enhancing team cognition in complex environments: The power of team performance diagnosis. *Aviation, space, and environmental medicine*, *78*(5), B77-B85.

Salen, K., Tekinbaş, K. S., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. MIT press.

Sarma, K. M., Carey, R. N., Kervick, A. A., & Bimpeh, Y. (2013). Psychological factors associated with indices of risky, reckless and cautious driving in a national sample of drivers in the Republic of Ireland. *Accident Analysis & Prevention*, 50, 1226-1235. https://doi.org/10.1016/j.aap.2012.09.020

Savadori, L., Van Swol, L. M., & Sniezek, J. A. (2001). Information sampling and Confidence within groups and judge advisor systems. *Communication Research*, *28*(6), 737-771. https://doi.org/10.1177/009365001028006002

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). New York, NY: Guilford Press.

Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology*, *9*(6), 523-538. https://doi.org/10.1002/acp.2350090605

Schneider, F., Calapai, A., Mundry, R., Báez-Mendoza, R., Gail, A., Kagan, I., & Treue, S. (2024). Confidence over competence: Real-time integration of social information in human continuous perceptual decision-making. *bioRxiv*, 2024-08. https://doi.org/10.1101/2024.08.19.608609

Schuldt, J. P., Chabris, C. F., Woolley, A. W., & Hackman, J. R. (2017). Confidence in dyadic decision making: The role of individual differences. *Journal of Behavioral Decision Making*, *30*(2), 168–180. https://doi.org/10.1002/bdm.1927

Selenta, C., & Lord, R. G. (2005). Development of the levels of self-concept scale: measuring the individual, re- lational, and collective levels. Working paper, University of Akron.

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: the links between language, performance, error, and workload. *Human Performance in Extreme Environments*, 5, 63-68. http://docs.lib.purdue.edu/jhpee/vol5/iss1/6

Shanks, D., Brydges, R., den Brok, W., Nair, P., & Hatala, R. (2013). Are two heads better than one? Comparing dyad and self-regulated learning in simulation training. Medical education, 47(12), 1215-1222. https://doi.org/10.1111/medu.12284

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 43-58). New York: NY, Springer.

Shipstead, Z., Lindsey, D. R., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, 72, 116–141. https://doi.org/10.1016/j.jml.2014.01.004

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, *4*(3-4), 181-201. https://doi.org/10.1016/0004-3702(73)90011-8

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80. https://doi.org/10.1177/1745691613514755

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and Confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*(1), 1–28. https://doi.org/10.1016/0749-5978(89)90055-1

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65*(2), 117-137. https://doi.org/10.1006/obhd.1996.0011

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–292. https://doi.org/10.2307/1412107

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). *Latent profile analysis: A review and "how to" guide of its application within vocational behavior research*. Journal of Vocational Behavior, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Stachowski, A. A., Kaplan, S. A., & Waller, M. J. (2009). The benefits of flexible team interaction during crises. *Journal of Applied Psychology*, *94*(6), 1536. https://doi.org/10.1037/a0016903

Stankov, L. (1997). *The Gf/Gc Quickie Test Battery.* Unpublished test battery available from the School of Psychology, University of Sydney.

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*(6), 971-986. https://doi.org/10.1016/S0191-8869(96)00130-4

Stankov, L., Kleitman, S., & Jackson, S. A. (2014). Measures of the trait of Confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs* (pp. 158–189). London, UK: Academic Press.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012a). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747-758. https://doi.org/10.1016/j.lindif.2012.05.013

Stankov, L., Pallier, G., Danthiir, V., & Morony, S. (2012b). Perceptual underconfidence: A conceptual illusion? *European Journal of Psychological Assessment, 28*(3), 190-200. https://doi.org/10.1027/1015-5759/a000126

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, *48*(6), 1467. https://psycnet.apa.org/doi/10.1037/0022-3514.48.6.1467

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of personality and social psychology*, *53*(1), 81.

Steiner, I. D. (1972). *Group process and productivity* (pp. 393-422). New York: Academic press.

Stevens, J. P. (2002). Applied multivariate statistics for the social sciences (4th ed.). Mahwah, NJ: Erlbaum.

Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management*, *32*(1), 29-55. https://doi.org/10.1177/0149206305277792

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors*, *41*(1), 61-71. https://doi.org/10.1518/001872099779577273

Sutcliffe, K. M., Lewton, E., & Rosenthal, M. M. (2004). Communication failures: An insidious contributor to medical mishaps. *Academic Medicine*, 79, 186–194. https://doi.org/10.1097/00001888-200402000-00019

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Boston, MA: Pearson.

Tazelaar, M. J. A., Van Lange, P. A. M., & Ouwerkerk, J. W. (2004). How to Cope With "Noise" in Social Dilemmas: The Benefits of Communication. *Journal of Personality and Social Psychology, 87*(6), 845–859. https://doi.org/10.1037/0022-3514.87.6.845

Tindale, R. S. (1989). Group vs individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes*, *44*(3), 454–473. https://doi.org/10.1016/0749-5978(89)90019-8

Tindale, R., & Kameda, T. (2000). "Social Sharedness" as a Unifying Theme for Information Processing in Groups. *Group Processes & Intergroup Relations*, *3*(2), 123–140. https://doi.org/10.1177/1368430200003002002

Tolsgaard, M. G., Madsen, M. E., Ringsted, C., Oxlund, B. S., Oldenburg, A., Sorensen, J. L., ... & Tabor, A. (2015). The effect of dyad versus individual simulation-based ultrasound training on skills transfer. Medical education, 49(3), 286-295. https://doi.org/10.1111/medu.12624

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168. https://doi.org/10.1080/13546783.2013.844729

Treen, E., Atanasova, C., Pitt, L., & Johnson, M. (2016). Evidence from a large sample on the effects of group size and decision-making time on performance in a marketing simulation game. *Journal of Marketing Education*, *38*(2), 130-137. https://doi.org/10.1177/0273475316653433

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Hiel, A., Vanneste, S., & De Cremer, D. (2008). Why did they claim too much? The role of causal attributions in explaining level of cooperation in commons and anticommons dilemmas. *Journal of Applied Social Psychology*, 38, 173–197. https://doi.org/10.1111/j.1559-1816.2008.00301.x

Vidulich, M.A., & Bortolussi, M.R. (1988). A dissociation of objective and subjective workload measures in assessing the impact of speech controls in advanced helicopters. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 1471–1475). Santa Monica, CA: Human Factors Society.

Voracek, M., and Dressler, S. G. (2006). Lack of correlation between digit ratio (2D:4D) and Baron-Cohen's "Reading the Mind in the Eyes" test, empathy, systemising, and autism-spectrum quotients in a general population sample. *Personality and Individual Differences*, 41, 1481–1491. https://doi.org/10.1016/j.paid.2006.06.009

Vroom, V. H., & Yetton, P. W. (1973). *Leadership and Decision-making* (Vol. 110). University of Pittsburgh Pre.

Wildman, J. L., Salas, E., & Scott, C. P. (2014). Measuring cognition in teams: A cross-domain review. *Human factors*, *56*(5), 911-941. https://doi.org/10.1177/0018720813515907

Wilhelm, O., Hildebrandt, A., Manske, K., Schacht, A., & Sommer, W. (2014). Test battery for measuring the perception and recognition of facial expressions of emotion. *Frontiers in Psychology*, *5*, 404. https://doi.org/10.3389/fpsyg.2014.00404

Williams, W. M., & Sternberg, R. J. (1988). Group intelligence: Why some groups are better than others. *Intelligence*, *12*(4), 351-377. https://doi.org/10.1016/0160-2896(88)90002-5

Wilson, J. M., Straus, S. G., & McEvily, B. (2006). All in due time: The development of trust in computer-mediated and face-to-face teams. *Organizational Behavior and Human Decision Processes*, 99, 16-33. https://doi.org/10.1016/j.obhdp.2005.08.001

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), Metacognition in educational theory and practice (pp. 277-304). Hillsdale, NJ: Lawrence Erlbaum

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. https://doi.org/10.1126/science.1193147

Xie, X., Yu, Y., Chen, X., & Chen, X. (2006). The Measurement of Cooperative and Competitive Personality. *Acta Psychologica Sinica, 38*(1), 116–125.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, *69*, 237–249. https://doi.org/10.1006/obhd.1997.2685

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology*, *33*(4), 345-366. https://doi.org/10.1006/jesp.1997.1326

Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., ... & Herzog, S. M. (2024). Human-AI collectives produce the most accurate differential diagnoses. *arXiv preprint arXiv:2406.14981*.

Appendix A: Supplementary Material for Chapter 2 (Study 1)

Participant Exclusions

A total of 22 participants (12 individuals and 5 dyads) were excluded from our analyses. Eight participants (6 individuals and 1 dyad) were excluded because they did not complete the driving simulation. Six participants (3 dyads) were excluded due to a computer error that caused one dyad member to observe a marginally different simulated environment from their teammate which rendered the navigator's instructions inaccurate. Three participants (3 individuals) were excluded because they did not follow instructions during the driving simulation. Five participants (3 individuals and 1 dyad) were excluded because they were identified as outliers (~4+ SDs from mean) on the dependent variables by inspecting scatterplots and computing Mahalanobis distance and Cook's *D* for each participant (Stevens, 2002).

Frequency Distributions for Each Condition

Figure A1. *Frequency Distributions for Individual Collisions (A) and Speed (B) and Dyad Collisions (C) and Speed (D) During the Normal Condition*

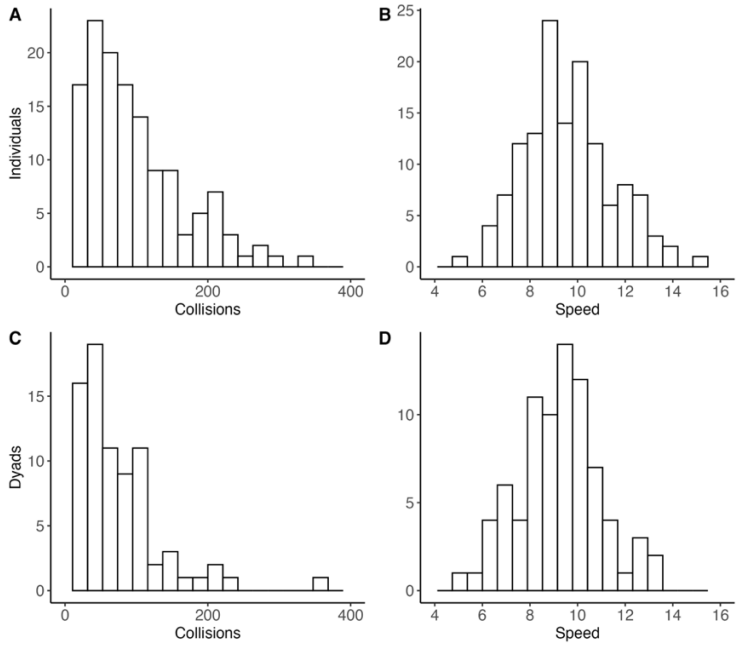
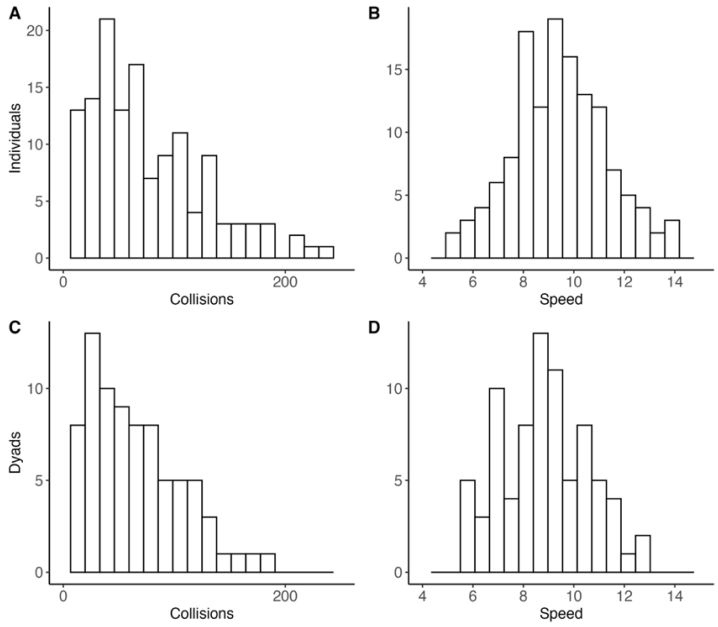


Figure A2. *Frequency Distributions for Individual Collisions (A) and Speed (B) and Dyad Collisions (C) and Speed (D) During the Fog Condition*



Individual Difference Measures

Descriptive statistics and internal consistency for the individual difference measures used to assess the known key covariates are presented in Table A1 for individuals and dyads and Table A2 for drivers and navigators within dyads.

Table A1. *Descriptive Statistics for the Psychometric Measures and t-tests Comparing Individuals and Dyads*

	Individuals			Dyads			
	IC	Mean	SD	IC	Mean	SD	*t*
Executive Functions							
Switching							
Repeat time	.93	983	367	.89	933	281	1.13
Switch time	.95	1067	387	.88	1042	306	0.52
Repeat errors	.67	1.62	2.25	.57	1.67	1.80	-0.18
Switch errors	.54	2.41	2.43	.41	2.12	2.13	0.96
Inhibition							
Congruent time	.69	463.09	46.07	.97	458.32	49.62	0.79
Incongruent time	.82	531.51	66.78	.84	530.31	62.61	0.18
Congruent errors	.39	1.41	1.92	.47	0.92	1.31	1.99*
Incongruent errors	.63	1.66	2.06	.55	1.52	1.73	0.50
Working Memory							
Accuracy	.71	44.78	18.64	.71	43.04	18.69	0.72
Fluid Intelligence							

Accuracy	.81	53.06	21.33	.82	57.05	21.38	-1.49
Confidence							
Cognitive Confidence	.92	61.07	18.21	.92	61.97	18.05	-0.40
Simulation Confidence	-	2.88	1.22	-	3.70	0.81	-5.34***
Personality							
Agreeableness	.59	3.17	0.60	.68	3.90	0.67	-9.48***
Conscientiousness	.60	3.08	0.60	.64	3.29	0.72	-2.63**
Extraversion	.79	2.91	0.52	.79	3.12	0.81	-2.56*
Intellect	.55	2.84	0.61	.58	3.59	0.65	-9.26***
Neuroticism	.47	3.02	0.50	.63	2.88	0.70	1.96

Note. IC = Internal Consistency. Internal Consistency estimates were computed using McDonald's Omega for all variables except switching, inhibition, Bias, and discrimination. For these variables we correlated scores on the odd items with scores on the even items then adjusted the correlation coefficient using the Spearman-Brown formula (Guilford, 1954; Stankov & Crawford, 1996).

$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$

Individuals and dyads significantly differed on simulation confidence and each of the personality facets except Neuroticism. Individuals were lower on simulation confidence, Agreeableness, Conscientiousness, Extraversion, and Intellect compared with dyads. Within dyads, drivers were significantly lower on simulation confidence and significantly faster than navigators on congruent time and incongruent time. No other significant differences were observed.

All reliability estimates were acceptable for research purposes except repeat errors (dyad a = .57 and driver a = .29), switch errors (all levels, a = .28 - .54), congruent errors (all levels, a = .39 - .54), incongruent errors (dyad a = .55 and driver a = .44), conscientiousness (codriver a = .56), intellect (individual = .55, dyad a = .58, driver a = .53), and neuroticism (individual a = .47) which ranged from low to poor. The repeat errors, switch errors, congruent errors, and incongruent errors variables consistently demonstrated poor internal consistency thus they were removed from subsequent analyses. These four variables assessed inhibitory control and cognitive flexibility. We had two different metrics for each of these constructs: errors and response time. The response time measures demonstrated excellent reliability (ranging from ω_t = .88 - .95), thus, they remained in the study for our analyses.

Reliability estimates for the personality measures ranged between ω_t = .47 - .79 for individuals and ω_t = .58 - .79 for dyads. Some of the reliability estimates for individuals were low, however, we only used the dyad measures as control variables to examine hypotheses related to aims 2 and 3. Overall, these estimates were consistent with previous literature using this brief instrument (Blanchard et al., 2020; Jackson et al., 2016; Jackson et al., 2017).

Table A2. *Descriptive Statistics for the Psychometric Measures and t-tests Comparing Drivers and Navigators*

	Driver			Navigator			
	IC	Mean	SD	IC	Mean	SD	*t*
Executive Functions							
Switching							
Repeat time	.87	941	258	.90	926	303	0.33
Switch time	.87	1040	291	.88	1044	322	-0.11
Repeat errors	.29	1.69	1.59	.69	1.65	2.01	0.18
Switch errors	.28	2.08	1.84	.49	2.16	2.39	-0.27
Inhibition							
Congruent time	.98	446.41	47.43	.95	470.39	49.16	-2.82**
Incongruent time	.87	515.76	63.06	.80	545.05	58.96	-2.72**
Congruent errors	.54	0.95	1.41	.39	0.90	1.19	0.31
Incongruent errors	.44	1.49	1.54	.63	1.56	1.91	-0.22
Working Memory							
Accuracy	.73	41.67	19.46	.68	44.42	17.91	-0.98
Fluid Intelligence							
Accuracy	.82	56.39	21.08	.83	57.71	21.79	-0.37
Metacognition							
Cognitive Confidence	.93	62.01	18.09	.92	61.94	18.13	0.02
Simulation Confidence	-	3.28	1.10	-	4.09	1.24	-4.37***
Personality							
Agreeableness	.71	3.95	0.67	.64	3.85	0.68	0.91
Conscientiousness	.69	3.18	0.74	.56	3.40	0.68	-1.80
Extraversion	.83	3.10	0.85	.74	3.14	0.78	-0.39
Intellect	.53	3.58	0.65	.63	3.60	0.67	-0.23
Neuroticism	.64	2.92	0.73	.62	2.84	0.66	0.72

Note. IC = Internal Consistency. Internal Consistency estimates were computed using McDonald's Omega for all variables except switching, inhibition, Bias, and discrimination. For these variables we correlated scores on the odd items with scores on the even items then adjusted the correlation coefficient using the Spearman-Brown formula (Blanchard et al., 2020; Guilford, 1954; Stankov & Crawford, 1996).

***$p < .001$; **$p < .01$

Reduction of independent variables

To retain adequate power in our hierarchical regression analyses for hypotheses 2 and 3, we reduced the independent variables for dyads down to a smaller number of components using PCA. These extracted components were: Executive Function Time which was composed of the response time variables for repeat time, switch time, congruent time, and incongruent time; and Competence which was composed of fluid intelligence, confidence, and working memory accuracy. These PCAs and the correlations between all outcome variables, dyad composition measures, and control variables are presented below.

Table A3

Intercorrelations and PCA Results for Executive Function Time Variables

	Pearson *r* correlations			Component loadings	
	2	3	4	1	h^2
1. Switch time	.89	.45	.34	**.84**	.70
2. Repeat time		.40	.29	**.81**	.65
3. Congruent time			.84	**.83**	.69
4. Incongruent time				**.75**	.57

Note. D=Driver; N=Navigator; EF variables = Executive Function variables; Component loadings >.30 are in bold. h^2= communality.
$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$

First, we extracted latent component(s) of Executive Function time. The correlations between these variables and a summary of the results of the PCA are presented in Table A3.

A pattern of small to large positive correlations was evident between all variables. We conducted a PCA (with Promax rotation) on repeat time, switch time, congruent time, and incongruent time. Inspection of scree plots, the Kaiser criterion, and Horn's Parallel Analysis (with 5000 iterations: Horn, 1965) suggested a two-component solution. However, to reduce the number of variables to include in the final model we extracted a single component which explained 65% of the common variance. All time variables loaded positively on this component which was named EF Time.

Next, we extracted latent component(s) of Competence. The correlations between these variables and a summary of the results of the PCA are presented in Table A4.

Table A4. *Intercorrelations and PCA Results for Competence Variables*

	Pearson *r* correlations		Component loadings	
	2	3	1	h^2
1. Fluid intelligence	0.40	0.35	**.72**	.52
2. Confidence	1	0.20	**.82**	.67
3. Working memory		1	**.67**	.45

Note. D=Driver; N=Navigator; Component loadings >.30 are in bold. h^2= communality.
$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$

A pattern of small to moderate positive correlations was evident between all variables. We conducted a PCA (with Promax rotation) on fluid intelligence, confidence, and working memory accuracy. Inspection of scree plots, the Kaiser criterion, and Horn's Parallel Analysis (with 5000 iterations: Horn, 1965) suggested a one-component solution. We extracted a single component which explained 55% of the common variance. All time variables loaded positively on this component which was named Competence.

The EF time and Competence components were included as control variables in the hierarchical regression analyses related to hypotheses 2 and 3.

Appendix B: Supplementary Material for Chapter 3 (Study 2)

Descriptive Statistics for Individual Difference and Communication Variables

Individual differences. The descriptive statistics and internal consistency estimates for the individual difference measures are reported in Table B1.

Table B1. *Descriptive Statistics and Internal Consistency Estimates for Individual Differences Measures (N=105)*

	N	ω_t	Mean	SD
Social sensitivity	103	.71	57.28	9.66
Working memory accuracy	105	.77	47.71	16.16
Agreeableness	104	.71	3.94	0.50
Conscientiousness	104	.73	3.20	0.64
Extraversion	104	.74	2.94	0.62
Intellect	104	.64	3.68	0.55
Neuroticism	104	.62	2.95	0.51

Note: ω_t = Internal consistency measured using Omega total

The means and standard deviations for social sensitivity, working memory accuracy, and personality were comparable with other studies that have used the same measures on an undergraduate population (Jackson et al., 2017; Law et al., 2018; Law et al., 2022). Internal consistency estimates ranged from acceptable (.66) to excellent (.89) for all measures.

Communication measures. Table B2 describes the descriptive statistics for the communication metrics captured during each test and overall. These variables were based on frequency counts thus, no estimate of internal consistency could be made.

Table B2. *Descriptive statistics for communication measures for each test and overall (N=101)*

Test	Items	Number of talking turns				Duration of discussion			
		Total		Equality		Total (seconds)		Equality	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
ADR	10	138.88	58.00	23.81	22.32	370.89	231.29	79.03	72.29
CRT	7	126.11	44.45	18.71	19.50	351.15	169.46	72.56	61.24
RAPM	18	348.81	153.21	41.56	40.01	1076.23	738.57	171.73	159.37
Overall		613.80	204.93	71.23	73.80	1798.26	980.05	279.58	270.03

Note. ADR = Applying Decision Rules, CRT = Cognitive Reflection Test; RAPM = Raven's Advanced Progressive Matrices.

The CRT had the fewest number of speaking turns, the shortest duration of discussion, the greatest equality of turn taking, and the greatest equality of duration, on average. Raven's Advanced Progressive Matrices, on the other hand, had the highest scores on each variable. This was not surprising given that it had the largest number and, arguably, the most complex items.

Additional Results for CFA Without Geography Test

Additional analyses were conducted to determine the best fitting CFA model. Table B3 reports the modification indices for the relationship between accuracy and confidence within each test for each model tested. Scores below ten are optimal. Tables B4, B5, and B6 report the standardised residual covariances for the models tested. Values greater than ±2.00 suggest poor model fit. The modification indices and standardised residual covariances indicate that the best fitting model was the modified two factor model with error terms correlated for RAPM and CRT.

Table B3. *Modification Indices for the Relationship Between Accuracy and Confidence for each test for the Unmodified Two Factor Model*

Model	U2	M2R	M2RC
RAPM	35.76	-	-
CRT	15.73	10.02	-
ADR	6.56	8.68	3.59

Note. U2 = Unmodified two factor model. M2R = Modified two factor model with the error terms for RAPM correlated. M2RC = Modified two factor model with the error terms for RAPM and CRT correlated. RAPM = Raven's Advanced Progressive Matrices. CRT = Cognitive Reflection Test. ADR = Applying Decision Rules.

Table B4. *Standardised Residual Covariances for the Unmodified Two Factor Model*

	ADR Acc	CRT Acc	RAPM Acc	ADR Conf	CRT Conf	RAPM Conf
ADR Acc	0.07					
CRT Acc	0.48	0.00				
RAPM Acc	-0.02	-0.31	0.00			
ADR Conf	1.16	-2.36	-1.26	-0.15		
CRT Conf	-0.90	1.48	-0.91	0.70	0.00	
RAPM Conf	-0.74	-0.74	3.83	1.08	-1.20	0.08

Note. RAPM = Raven's Advanced Progressive Matrices. CRT = Cognitive Reflection Test. ADR = Applying Decision Rules. Acc = Accuracy. Conf = Confidence.

Table B5. *Standardised Residual Covariances for the Modified Two Factor Model with the Error Terms for RAPM Correlated*

	ADR Acc	CRT Acc	RAPM Acc	ADR Conf	CRT Conf	RAPM Conf
ADR Acc	-0.03					
CRT Acc	-0.60	0.00				
RAPM Acc	0.17	-0.54	-0.68			
ADR Conf	1.51	-2.04	-0.51	-0.13		
CRT Conf	-0.61	2.20	-0.25	-0.79	0.00	
RAPM Conf	-0.33	-0.50	-0.74	1.32	-1.00	-0.43

Table B6. *Standardised Residual Covariances for the Modified Two Factor Model with the Error Terms for RAPM and CRT Correlated*

	ADR Acc	CRT Acc	RAPM Acc	ADR Conf	CRT Conf	RAPM Conf
ADR Acc	0.05					
CRT Acc	-0.29	0.48				
RAPM Acc	-0.91	1.63	0.21			
ADR Conf	0.63	-0.93	-0.74	-0.14		
CRT Conf	-0.41	0.19	1.52	-0.72	-0.12	
RAPM Conf	-0.77	1.14	0.29	-0.33	1.54	0.21

CFA Results with Geography Test Included

We fitted the same CFA models that are reported in the main paper with the addition of accuracy and confidence on the geography test (GT). The results are presented in Tables B7, B8, and B9. A modified two-factor model had the best fit for collective accuracy and confidence (model 3^{d}). In this model, the error terms of accuracy and confidence were correlated within the same test for RAPM and CRT. The fit indices for this modified two-factor model were excellent: $R^2 = .47$; $\chi^2/df = 1.22$; Goodness of Fit Index (GFI) = 0.99; Tucker-Lewis Index (TLI) = 0.98; Comparative Fit Index (CFI) = 0.99; Root Mean Square Error of Approximation (RMSEA) = 0.05 (CI = .00-.10). The results of this CFA model are displayed in Table B7. Accuracy from ADR, CRT, and RAPM load well onto the CI factor, however, GT accuracy loads poorly (.21, $p > .05$) and has a low communality (.04). Together, these values indicate that GT accuracy shares little variance with the underlying CI factor,

suggesting it does not integrate well into the model. Thus, it was removed from the model and excluded from the analyses.

Table B7. *Summary of Fit Indices Evaluating Different Models of Intelligence and Confidence for Dyads Using Maximum Likelihood CFA (N = 105)*

Model	Fit Statistics									
	R^2	χ^2	*df*	X^2 / *df*	χ^2 diff	GFI	TLI	CFI	RMSEA (90% CInt)	AIC
One-factor[a]	.42	94.65	20	4.73	-	0.99	.71	.79	.19 (.15-.23)	6619
Two-factor 1[b]	.47	78.94	19	4.15	15.71***	0.99	.75	.83	.17 (.13-.21)	6605
Two-factor 2[c]	.48	36.28	18	2.02	42.66***	0.99	0.92	0.95	.10 (.05-.14)	6564
Two-factor 3[d]	.47	20.66	17	1.22	15.62***	0.99	0.98	0.99	.05 (.00-.10)	6550

Note. GFI = Goodness-of-fit index; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; CInt = Confidence Interval; AIC = Akaike Information Criterion. The accepted model is in bold.

[a]One-factor model consisted of one broad first order Cognitive factor defined by all the measures employed in the study without any modifications to the model.

[b]Two-factor model consisted of an intelligence factor (defined by all accuracy measures) and a confidence factor (defined by all confidence measures) without any modifications to the model.

[c]Two-factor model (intelligence and confidence factors) where error terms of the corresponding accuracy and confidence scores from RAPM were correlated.

[d]Two-factor model (intelligence and confidence factors) where error terms of the corresponding accuracy and confidence scores from RAPM and CRT were correlated.

$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$

Table B8. *Summary of Standardised Regression Weights, Communalities, and Correlations from a CFA Using Dyadic Variables and Including (N = 105)*

Measures	Intelligence	Confidence	h^2
ADR accuracy	.68		.46
CRT accuracy	.73		.54
RAPM accuracy	.63		.40
GT accuracy	.21		.04
ADR confidence		.88	.78
CRT confidence		.84	.71
RAPM confidence		.69	.48
GT confidence		.59	.35
Factor intercorrelations			
Intelligence	1	.78	
Confidence		1	

Note. All loadings and the factor intercorrelation were significant with $p < .001$ except for GT accuracy which did not reach significance.

Table B9. *Standardised Residual Covariances for the Modified Two Factor Model with the Error Terms for RAPM and CRT Correlated*

	ADR Acc	CRT Acc	RAPM Acc	GK Acc	ADR Conf	CRT Conf	RAPM Conf	GK Conf
ADR Acc	0.04							
CRT Acc	0.21	0.62						
RAPM Acc	-0.54	1.79	0.32					
GT Acc	0.3	-0.8	-0.63	0.00				
ADR Conf	0.78	-0.25	-0.32	0.37	-0.08			
CRT Conf	-0.12	0.44	1.73	-0.58	-0.55	-0.02		
RAPM Conf	-0.59	1.38	0.45	-0.71	-0.28	1.38	0.31	
GT Conf	-2.11	-1.86	-1.34	0.99	-0.05	0.12	0.54	0.01

Hierarchical Regression Models Fit on Data Without Imputation

Table B10 presents the results of the hierarchical regression analyses fit using data without imputation.

Table B10. *Results of Hierarchical Regression Analyses Using Individual Variables to Predict Collective Intelligence and Confidence*

	Collective Intelligence			Collective Confidence		
	Block			Block		
	1	2	3	1	2	3
Predictor	β	β	β	β	β	β
Mixed gender dyads	-.15	-.25	-.17	-.32	-.33	-.32
Female dyads	-.84***	-.89***	-.57*	-1.12***	-1.10***	-.69**
Social sensitivity	.39***	.34***	.22**	.28**	.19*	.12
Inequality of turn taking	.08	.06	.08	.13	.09	.11
WM accuracy	-	.24*	.10	-	.36***	.23**
Agreeableness	-	-.01	.05	-	.14	.17*
Conscientiousness	-	-.00	-.04	-	.08	-.01
Extraversion	-	-.21*	-.10	-	-.02	.06
Intellect	-	.09	-.09	-	-.03	-.22**
Neuroticism	-	.09	-.02	-	.06	.01
Intelligence	-	-	.32**	-	-	-.03
Confidence	-	-	.33***	-	-	.63***
R	.47	.59	.77	.49	.61	.80
R^2	.22	.35	.59	.24	.37	.64
ΔR^2	.22***	.13*	.24***	.24***	.13*	.27***

Note. WM = Working Memory. β = standardised regression coefficient.
***$p < .001$; **$p < .01$; *$p < .05$

Summary of LPA Goodness of Fit Indices and Model Selection

LPA was performed for solutions with 2-6 classes (with 1 class as the default, see Table B11 on 6 predictor variables. These variables were individual intelligence and confidence, the extracted factors for CI and collective confidence, and individual and collective bias scores. Goodness of fit statistics were used to identify the number of latent classes (Henson et al., 2007; Marsh et al., 2009). Assessment of the indices and examination of the profiles within each model suggested a 3-Class solution was the best fitting model.

Table B11. *Goodness of Fit Statistics for All Latent Profile Analysis Models Tested (N = 105)*

Classes in the model	AIC	Adjusted BIC	BIC	Entropy	BLRT	LogLik
1	1806	1800	1838	1.00	-	-891
2	1603	1590	1669	0.94	229*	-776***
3	1560	1541	1661	0.96	69*	-742***
4	1560	1534	1695	0.96	26	-729***
5	1513	1480	1682	0.92	73*	-692***
6	1490	1451	1694	0.92	49*	-668***
Class	Class counts and proportions for the latent classes		Average latent class probabilities for most likely latent class membership (row) by latent class (column)			
	Counts	Proportions	Class 1	Class 2		
Class 1	64	.61	.99	.01		
Class 2	41	.39	.02	.98		
	Counts	Proportions	Class 1	Class 2	Class 3	
Class 1	57	.54	.99	.00	.01	
Class 2	15	.14	.00	.94	.06	
Class 3	33	.31	.05	.00	.95	
	Counts	Proportions	Class 1	Class 2	Class 3	Class 4
Class 1	52	.50	.95	.00	.00	.05
Class 2	15	.14	.00	.97	.00	.03
Class 3	5	.05	.00	.00	1.00	.00
Class 4	33	.31	.03	.01	.00	.96

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; BLRT = Bootstrap Likelihood Ratio Test; LogLik = Log Likelihood of the data, given the model. *p*-values of the chi-squared test between k and k-1 solutions.
$^{***}p < .001$; $^{*}p < .05$

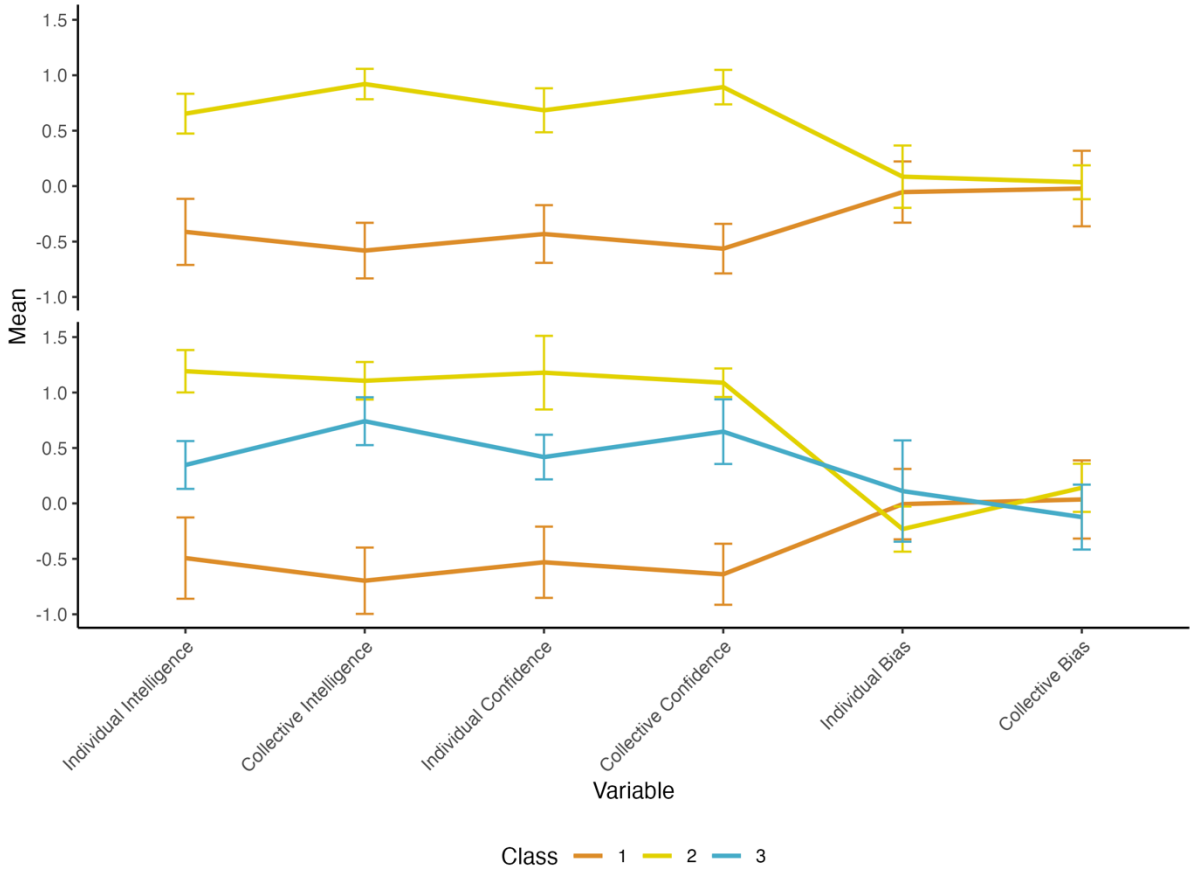
In particular, Akaike Information Criterion (AIC) values declined from the 1-Class to the 3-Class solution and then plateaued for the 4-Class solution. The Bootstrap Likelihood Ratio Test (BLRT) also suggested a 3-Class solution as it was a significant improvement on the 2-Class solution, but the 4-Class solution was not a significant improvement on the 3-

Class solution. All solutions had acceptable entropy values - entropy greater than .80 is considered high (Clark & Muthén, 2009) – and entropy was greatest for the 3- and 4-Class solutions (.96). Entropy was reverse coded so 1 = complete certainty of classification and 0 = complete uncertainty. The sample adjusted Bayesian Information Criterion (BIC) and Log-likelihood of the data, given the model (LogLik) indicated that each new model improved the fit – suggested by decreasing values for each subsequent solution and *p*-values for less than .05.

The 2-, 3-, and 4-Class solutions all appear to be viable candidates for selection, so we compared them on the proportion of class membership, the goodness of classification, and the interpretability of each class. The 2- and 3-Class models had an adequate proportion of members (> .10) in each profile, but the 4-Class model did not (Class 3 = .05). Next, we checked the goodness of classification by examining the average latent class probabilities for the most likely latent class membership. All were 90% or greater for the assigned class which is acceptable (See Table B11: high values on the diagonal and low values off the diagonal indicated goodness of classification).

To select a model from the 2- and 3-Class solutions we examined the interpretations of the profiles within each model (see Figure B1). Both models contained low and high CI profiles, but the 3-Class solution also included an amplified CI profile whose members had significantly greater CI than individual intelligence. This third profile identified a small number of participants that benefitted the most from working together in a dyad and provided additional insight into systematic outcomes for individuals paired to work together. We conducted the same analyses (described below) on both the 2- and 3-Class solutions and the pattern of results was consistent. Thus, we selected the 3-Class solution as the best fitting model because it provided additional information beyond the 2-Class solution. All subsequent analyses reported in the main text are based on this model.

Figure B1 . *Latent profile groups for 2- (A) and 3-Class (B) solutions. Error bars represent the standard error of the mean for each profile on each variable*



Appendix C: Supplementary Material for Chapter 4 (Study 3)

Development of the General-Knowledge Tests

Test development was conducted to construct three versions of a general knowledge test to administer in the main study under different communication conditions. These versions were designed to be matched on decision accuracy, decision confidence, and content domains.

Participants

Participants were 67 undergraduate psychology students (22 males, Mean age = 20.27, SD = 2.97) who completed the study for partial course credit. These participants were ineligible for participation in the later stages of this study.

Measures

The items for the general-knowledge test were collected from several previous studies (Brewer & Sampaio, 2012; Blanchard et al. 2020; Schuldt et al., 2017; Stankov, 1997) and constructed by members of the researcher team. 101 two alternative items were selected to cover a broad range of content areas: geography, art, music, film, history, science, and vocabulary. For example, *What does the word orthodox mean? Religious or Conventional** and *Who wrote the novel titled Brave New World? Aldous Huxley* or George Orwell* (* indicates the correct answer). After each item, participants were asked to provide a confidence rating ranging from 50% (guessing) to 100% (completely certain) for the correctness of their response. Of the 101 items, 22 were selected to form each of the 3 general-knowledge tests. By design, the three versions were matched on decision accuracy, decision confidence, and content domain.

Procedure

All participants completed the test development stage in a university computer lab. After providing consent, participants completed a demographic questionnaire then answered all general-knowledge items by themselves in a randomised order.

Test Development

Overall, mean decision accuracy was 64.49 (SD = 21.64) and mean decision confidence was 73.40 (SD = 20.41) for the general-knowledge items. We used item-level decision accuracy and decision confidence to construct 3 different versions of the general-knowledge test that had equivalent mean decision accuracy and decision confidence. Table C1 demonstrates the descriptive statistics for each version of the general-knowledge test that we constructed from the total pool of items. There were no significant differences on mean decision accuracy ($F = .04$, $p = .84$) or decision confidence ($F = .01$, $p = .91$) between the three matched versions of the general-knowledge tests. Omega total was used to measure internal consistency (McDonald, 1999). For decision accuracy, internal consistency was acceptable for exploratory research purposes for version 2 and 3 ($\omega_t = .56$ and .63, respectively) but was low for version 1 ($\omega_t = .47$). For decision confidence, internal consistency was good to excellent (ω_t ranging from .75 to .82).

Table C1. *Descriptive Statistics for the Matched Versions of the General Knowledge Test (N = 67)*

	Version			
	1	2	3	*F*
Decision Accuracy				
Mean	58.28 (0.20)	58.07 (0.20)	57.94 (.20)	0.04
ω_t	.47	.56	.63	
Decision Confidence				
Mean	71.64 (19.56)	71.58 (20.29)	71.70 (19.92)	0.01
ω_t	.81	.75	.82	

Note. ω_t = Omega total.

The Items for the Final Versions of the General Knowledge Tests Used in the Main Study

Table C2. *The General Knowledge Test Items for the Isolated Communication Condition*

Item	Question	Option 1	Option 2	Correct Response
1	What does the word duress mean?	Period of time	Compulsion	Compulsion
2	What does the word abjure mean?	Renounce	Arrest	Renounce
3	Who wrote the novel titled Frankenstein?	Mary Shelley	Jane Austen	Mary Shelley
4	Who was the first president of the United States?	Abraham Lincoln	George Washington	George Washington
5	What does the word gush mean?	Spurt	Cry	Spurt
6	Approximately how many Australian Aboriginal languages are there?	150	300	300
7	The Dalai Lama is from which country?	Tibet	India	Tibet
8	In which city would you find the most famous works by Antoni Gaudi?	New York	Barcelona	Barcelona
9	Which school of art did the painter Claude Monet come from?	Impressionism	Expressionism	Impressionism
10	What does the word unwary mean?	Tireless	Incautious	Incautious

Table C3. *The General Knowledge Test Items for the Passive Communication Condition*

Item	Question	Option 1	Option 2	Correct Response
1	Which is a stress hormone produced by the human body?	Serotonin	Cortisol	Cortisol
2	Andy Warhol designed an album cover for which band?	The Beatles	The Rolling Stones	The Rolling Stones
3	How many feature films has Quentin Tarantino directed?	7	9	9
4	Who wrote the novel titled The Handmaid's Tale?	John Steinbeck	Margaret Atwood	Margaret Atwood
5	What does the word perspire mean?	Struggle	Sweat	Sweat
6	In which year did the Berlin wall fall?	1989	1979	1989
7	Which planet is larger in size?	Neptune	Jupiter	Jupiter
8	The Persistence of Memory is a painting by which artist?	Salvador Dali	Henri Matisse	Salvador Dali
9	What does the word feign mean?	Pretend	Be cautious	Pretend
10	A deficiency of vitamin C causes which disorder?	Scurvy	Rickets	Scurvy

Table C4. *The General Knowledge Test Items for the Active Communication Condition*

Item	Question	Option 1	Option 2	Correct Response
1	Which continent has a larger area?	Antarctica	Europe	Antarctica
2	Which country was Pablo Picasso born in?	France	Spain	Spain
3	For which film did Cate Blanchett win an Oscar award for Best Actress?	Blue Jasmine	Carol	Blue Jasmine
4	Who wrote the novel titled Anna Karenina?	Ernest Hemmingway	Leo Tolstoy	Leo Tolstoy
5	When did Sydney host the Summer Olympic Games?	2004	2000	2000
6	What does the word dyschronometria mean?	Impaired ability to estimate amount of time passed	Impaired ability to maintain a line of thought	Impaired ability to estimate amount of time passed
7	Which continent has a larger area?	Europe	North America	North America
8	Who wrote the Australian novel titled Cloudstreet?	Tim Winton	Christos Tsiolkas	Tim Winton
9	What does the word orthodox mean?	Religious	Conventional	Conventional
10	Who painted the ceiling of the Sistine Chapel?	Michelangelo	Raphael	Michelangelo

Pre-screening Study

The pre-screening study was used to identify participants who were high-trait or low-trait confidence for inclusion in the main study. Our selection criteria aimed to recruit individuals who scored within ±1.50 standard deviations of the mean on cognitive ability and beyond ±0.50 standard deviations on trait confidence. The target sample size for the main study was 210 participants, consisting of 105 high-trait and 105 low-trait confidence individuals. These participants were paired into 35 dyads in each of the trait confidence categories: low-trait, mixed-trait, and high-trait.

Figure C1 displays the distribution of standardised trait confidence and cognitive ability scores from the screening study and highlights the subset of participants who completed the main study. As shown, most selected individuals fall within the intended ranges on trait confidence and cognitive ability. However, due to attrition, a small number of participants fall slightly outside these thresholds. One notable pattern is that high-trait

confidence participants in the main study tend to cluster towards the upper end of the cognitive ability target range more than low-trait confidence individuals. This imbalance necessitated statistical control for the effect of cognitive ability in the main study analyses. Figure C2 presents histograms for both variables of the distributions for participants recruited in the main study.

Participants

Participants were 1251 Australian university students (370 males; mean age = 20.89, SD = 5.01) who completed the study for either partial course credit or financial reimbursement. A total of 62 participants were excluded for non-genuine responses for RAPM or EAT. The final sample were 1189 participants (343 males; mean age = 20.91, SD = 5.09) who were screened for trait confidence and cognitive ability.

Figure C1. *Distribution of Trait Confidence and Cognitive Ability Scores in the Screening Study with Highlighted Main Study Participants (N = 1189)*

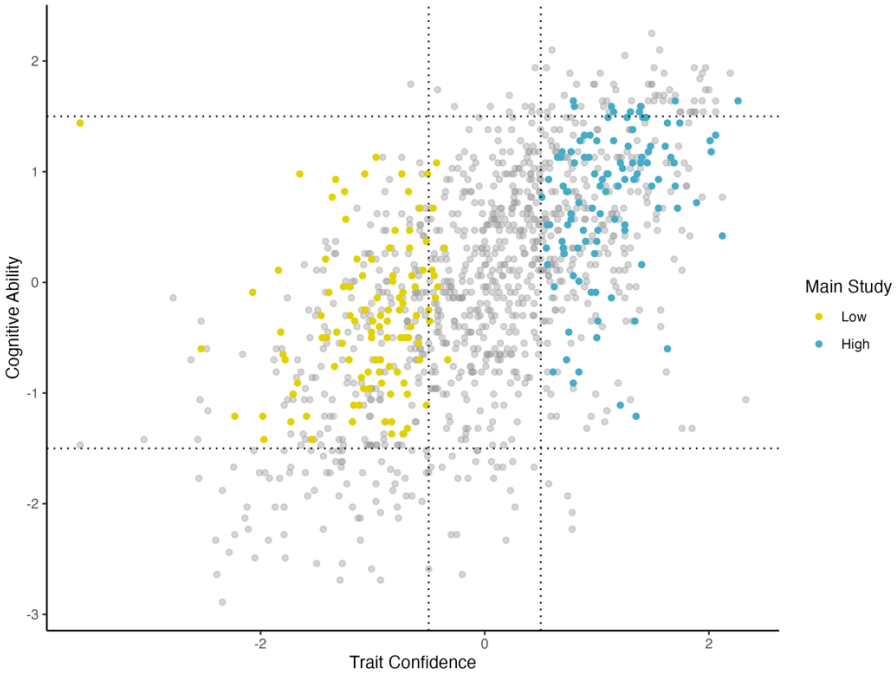
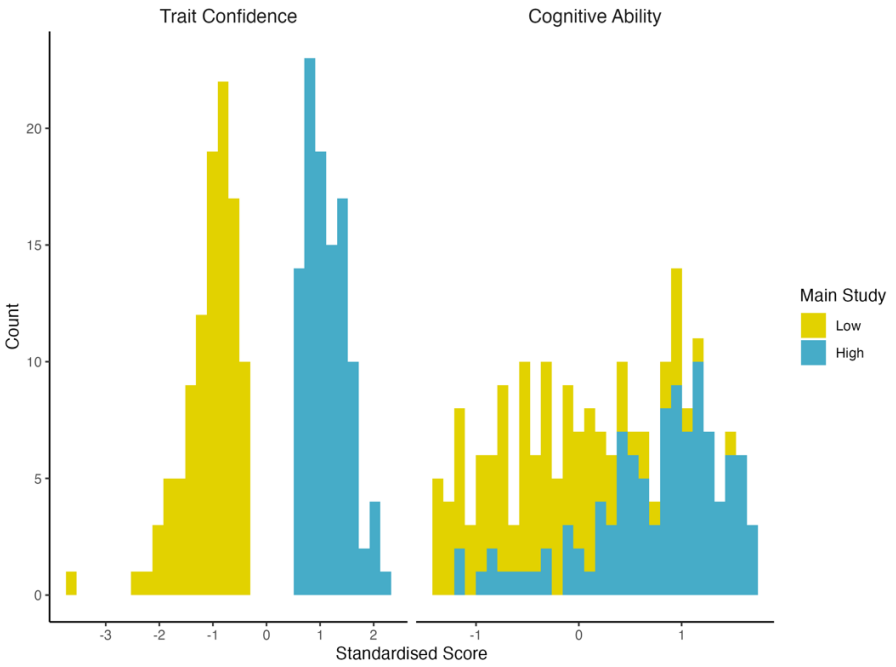


Figure C2. *Histograms of Trait Confidence and Cognitive Ability Scores in the Main Study (N = 210)*



Measures

Raven's Advanced Progressive Matrices (Raven, 1938-65): This test consists of 36 items, each featuring a 3x3 grid of abstract figures forming a horizontal and vertical pattern, with the bottom right figure missing. Participants select one of eight possible options to complete the matrix. Accuracy is a measure of Fluid Reasoning. The internal consistency has been shown to be excellent for accuracy, ranging from .80 to .81, and decision confidence, ranging from .90 to .92 (Blanchard et al., 2020; Blanchard et al., 2023). After responding to each item, participants rated their decision confidence in their answer on a scale from 12.5% (guessing) to 100% (completely certain). In the present study, participants completed a short 15-item version.

Mini-IPIP (Donnellan et al., 2006): This questionnaire presented participants with 20 statements about their personality, which they rated on a five-point scale ranging from "very inaccurate" (1) to "very accurate" (5). For example, one item asked participants to rate the accuracy of the statement "Am the life of the party." This scale assesses the Big Five personality traits and has been found to have acceptable internal consistency for Agreeableness (.70), Conscientiousness (.69), Extraversion (.77), Intellect (.65), and Neuroticism (.68).

Esoteric Analogies Test (EAT; Stankov, 1997): This measure involved participants completing 20 of the original 24 analogies. For each analogy, participants were presented with a pair of words and asked to select one of four options that reflected the same relationship with a target word. For example, *LOVE is to HATE as FRIEND is to: (1) LOVER, (2) PAL, (3) OBEY, (4) ENEMY**. Accuracy requires both reasoning skills and prior knowledge thus it is a mixed measure of Fluid Reasoning and Crystallised Intelligence. Prior research with Australian undergraduate samples reported acceptable internal consistency for

accuracy (ranging from .69 to .74) and excellent internal consistency for decision confidence (ranging from .88 to .94; Jackson et al., 2016; Law et al., 2022).

Procedure

All participants completed the 30-minute screening study remotely using their own device and internet connection. After providing consent, participants completed the tasks in the same order: demographic questionnaire, RAPM, mini-IPIP, and EAT. All measures were completed individually.

Individual Difference Measures from the Main Study

Motivational Traits Scale (Haesevoets et al., 2019): This scale consisted of 57 items that assessed 3 motivational traits associated with social decision-making: prosocial, proself, and fearful. The items were taken from scales that assess fairness (Van Hiel et al., 2008), altruism (Tazelaar et al., 2004), social welfare concerns (Haesevoets et al., 2018), concern for others (Selenta & Lord, 2005), greed (Krekels & Pandelaere, 2015), competitiveness (Xie et al., 2006), entitlement (Campbell et al., 2004), fear (Van Hiel et al., 2008) and risk aversion (Mandrik & Bao, 2005). For example, participants rated their agreement with statements like, *When I have to make a decision that also influences others I want to make a decision that leads to an equal outcome for everyone,* using a 7-point Likert scale, ranging from *Strongly Disagree (1)* to *Strongly Agree (7)*. The underlying scales have demonstrated good internal consistency with Cronbach's Alpha ranging from of .74 to .93.

Trust Scale (McAllister, 1995; Wilson et al., 2006): This scale uses 5 items to measure trust between members of small groups. It is appropriate for computer-mediated communication and McAllister's original scale was adapted by Wilson and colleagues for suitability with a student population. Participants rated their agreement with statements like *I can freely share my ideas and feelings in this group* using a 5-point Likert scale, ranging from *Strongly Disagree* (1) to *Strongly Agree* (5). Higher scores reflect greater trust withing a

group. The scale has been shown to possess excellent internal consistency for cognitive trust and affective trust which have Cronbach's Alpha values of .82 and .88 respectively.

Psychological Safety Scale (Edmondson, 1999): This scale uses 7 items to measure the extent to which individual group members feel safe to take interpersonal risks within their group. Participants rated their agreement with statements such as *If you make a mistake on this team, it is often held against you*, using a 5-point Likert scale from *Strongly Disagree* (1) to *Strongly Agree* (2). Higher scores reflect greater psychological safety within a group. The scale possess excellent internal consistency, with Cronbach's Alpha of .82 in the original study.

Empathy Quotient (Baron-Cohen & Wheelwright, 2004): The Empathy Quotient is a 60-item self-report questionnaire designed to measure empathy, which is the ability to understand and respond to the emotions and mental states of others. Of the 60 items, 40 assess empathy and 20 are control items. Participants responded to statements such as, *I can easily tell if someone is upset, even if they don't say anything,* using a 4-point Likert scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (4). Higher scores reflect greater empathy. The scale captures both cognitive empathy (understanding the mental states of others) and affective empathy (the capacity to respond emotionally to the feelings of others). Internal consistency has been shown to be excellent, with Cronbach's Alpha values around .92.

Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). This is a 36-item measure of emotion perception that assesses one's ability to infer emotional states from images of people's eyes. Woolley et al., (2010) used this measure to assess social sensitivity. Participants are presented with an image of someone's eyes and must quickly select the word (from four options) that best describes the thought or feeling expressed by the person in the image. Higher scores indicate greater ability at perceiving the emotions of others. In the

present study, participants completed a short 10-item version (Olderbak et al., 2015) which has been shown to possess good internal consistency, with Cronbach's Alpha of .73.

Behavioral Inhibition System/Behavioral Activation System Scales (BIS/BAS; Carver & White, 1994): The BIS/BAS consists of 24 items designed to measure individual differences in sensitivity to punishment (BIS) and reward (BAS). The BIS subscale assesses the degree to which participants experience behavioural inhibition in response to potential punishment or negative outcomes. In contrast, the BAS subscale captures the extent to which individuals are driven by rewards, divided into three components: Reward Responsiveness, Drive, and Fun Seeking. Participants rated statements such as *Criticism or scolding hurts me quite a bit* on a four-point scale, from *Very true for me* (1) to *Very false for me* (4). The internal consistency of the scales is good with Cronbach's Alpha scores of .74 for the BIS, .73 for the BAS Reward Responsiveness, .76 for the Bas Drive, and .66 for the BAS Fun Seeking subscales.

Risk Aversion (Holt & Laury, 2002). This behavioural measure comprised 10 items where participants made choices between two lottery options. Option A offered smaller, more stable payouts (lower risk), while Option B provided larger, more variable payouts (higher risk). For instance, in the first item, Option A gave a 1/10 chance of paying $2.00 and a 9/10 chance of paying $1.60, while Option B had a 1/10 chance of paying $3.85 and a 9/10 chance of paying only $0.10. Participants were required to choose which lottery they would prefer. Across items, the payout amounts for both options remained constant, but the probabilities of winning shifted incrementally with each successive item, moving by 1/10 with each step. By the final item, the odds reached 10/10 for both options, fully favouring the prize with the higher payout. This gradually increased the likelihood of the larger payout. The point at which a participant switches from selecting Option A to Option B provides an indicator of their risk aversion. For example, choosing Option B on the first item (when the probability of

the higher payout is just 1/10) indicates low risk aversion, whereas continuing to choose Option A on item 8 (where the probability of a high payout is 8/10) suggests very high risk aversion.

Communication Measures. Conversations between dyad members were recorded during the active communication condition. From these recordings, we calculated the number of speaking turns and the equality of turn-taking. Following Woolley et al. (2010), equality of turn-taking was measured by computing a standard deviation for the total number of speaking turns of a dyad's members. A zero-value indicated perfect equality, where both members contributed an equal number of turns, while higher values reflected increasing levels of inequality. For clarity, we referred to this measure as *inequality of turn-taking*.

Descriptive Statistics

Table C5 displays the descriptive statistics and internal consistency estimates for the additional individual difference variables. Internal consistency estimates ranged from acceptable (.64) to excellent (.91) for all psychological measures except social sensitivity which was low (.47).

The *inequality of* communication variables represent the similarity between dyad members for the number of speaking turns (inequality of turn-taking) and the number of words spoken (inequality of words spoken). A score of zero on either of the inequality variables indicates that dyad members had an identical number of speaking turns or words spoken and higher values indicated greater inequality.

The *total* communication variables represent the total number of speaking turns (total talking turns) and the total number of words spoken (total words spoken) for both dyad members. Internal consistency estimates were acceptable for inequality of turn-taking (.59) and good for the other communication variables, ranging from .74 to .83.

Table C5. *Descriptive Statistics and Internal Consistency Estimates for Individual Difference Variables for Each Trait Confidence Condition (N = 105)*

		Trait Confidence			
		Low	Mixed	High	
Variable	IC	Mean (SD)	Mean (SD)	Mean (SD)	$F_{2, 102}$
Social Sensitivity	.47	76.19 (20.96)	79.05 (17.88)	72.38 (27.93)	1.53
EQ	.89	44.94 (12.35)	43.51 (11.77)	42.43 (11.55)	0.79
BIS Total	.82	22.50 (3.64)	21.84 (4.49)	21.53 (3.33)	1.16
BAS Drive	.75	10.74 (2.32)	11.29 (2.11)	10.80 (2.35)	1.22
BAS Fun	.70	12.03 (2.13)	11.86 (2.35)	12.24 (2.10)	0.54
BAS Reward	.70	17.43 (1.77)	17.03 (2.31)	17.20 (2.15)	0.65
Proself Factor	-	-0.01 (0.46)	0.03 (0.44)	-0.02 (0.47)	0.22
Prosocial Factor	-	0.08 (0.58)	-0.06 (0.55)	-0.03 (0.50)	1.29
Fearful Factor	-	0.06 (0.50)	0.00 (0.61)	-0.07 (0.56)	0.97
Risk Aversion	.76	4.87 (2.31)	4.30 (2.06)	4.50 (2.14)	1.25
Psychological Safety	.64	5.58 (0.79)	5.74 (0.65)	5.64 (0.67)	0.88
Trust	.72	4.16 (0.51)	4.15 (0.50)	4.05 (0.49)	0.95
Inequality Turn Taking	.59	2.48 (2.20)	2.91 (3.71)	2.45 (1.96)	0.30
Inequality Words Spoken	.74	123.10 (115.08)	114.03 (97.64)	92.80 (74.79)	0.88
Total Talking Turns	.81	109.40 (30.99)	105.37 (26.32)	110.53 (32.06)	0.29
Total Words Spoken	.83	880.14 (403.23)	772.80 (292.11)	835.24 (356.44)	0.81

Selecting the Best Fitting Model for Decision Accuracy and Decision Confidence

We compared 8 models to select the best fitting model for decision accuracy and decision confidence. All models included the same fixed effects structure that included three-way interactions for grouping (individual vs dyad), communication type (isolated vs passive vs active), and trait confidence (low-trait vs mixed-trait vs high-trait), along with covariates for two cognitive abilities (EAT accuracy and RAPM accuracy). The models differed only in their random effects structure.

For each outcome, the models compared were: 1) linear regression with fixed effects only (no random effects were included); 2) random intercepts for each individual; 3) random intercepts for each dyad; 4) random intercepts for both individuals and dyads; and 5) random intercepts and slopes for communication type across individuals; 6) random intercepts and slopes for communication type across dyads; 7) random intercepts and slopes for communication type across individuals and random intercepts for dyads; 8) random intercepts and slopes for communication type at both the individual and dyad levels. We compared the

models on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood values. Likelihood ratio tests (LRTs) compared nested models relative to simpler alternatives, with statistical significance indicated where relevant. See Table C6 for the model comparisons.

Table C6. *Comparison of the Models for Each Outcome*

	Decision Accuracy			Decision Confidence		
Model	AIC	BIC	LRT	AIC	BIC	LRT
1	10542	10650	-5250	9300	9408	-4629
2	10328	10441	-5142***	8470	8583	-4213***
3	10358	10471	-5157	8806	8919	-4381
4	10312	10431	-5133***	8456	8574	-4205***
5	10128	10267	-5037***	8058	8197	-4002***
6	10290	10428	-5118	8794	8933	-4370
7	10107	10250	-5025***	8035	8179	-3990***
8	10079	10249	-5007***	8023	8192	-3978***

Model 1: no random effects
Model 2: (1 | individual) and Model 3: (1 | dyad)
Model 4: (1 | individual) + (1 | dyad) and Model 5: (1 + communication | individual)
Model 6: (1 + communication | dyad) and Model 7: (1 + communication | individual) + (1 | dyad)
Model 8: (1 + communication | individual) + (1 + communication | dyad); *** $p < .001$

For both outcomes, Model 8, which included random intercepts and slopes for communication at both the individual and dyad levels, provided the best fit based on AIC, BIC, and log-likelihood, with significant improvements over all simpler models. Despite increased model complexity, Model 8 did not exhibit convergence or singularity issues and was therefore retained as the final model for both decision accuracy and decision confidence analyses.

Additional Results for the Final Models for Decision Accuracy and Decision Confidence

Figure C3. *Mean Scores on Each Outcome for the Conditions*

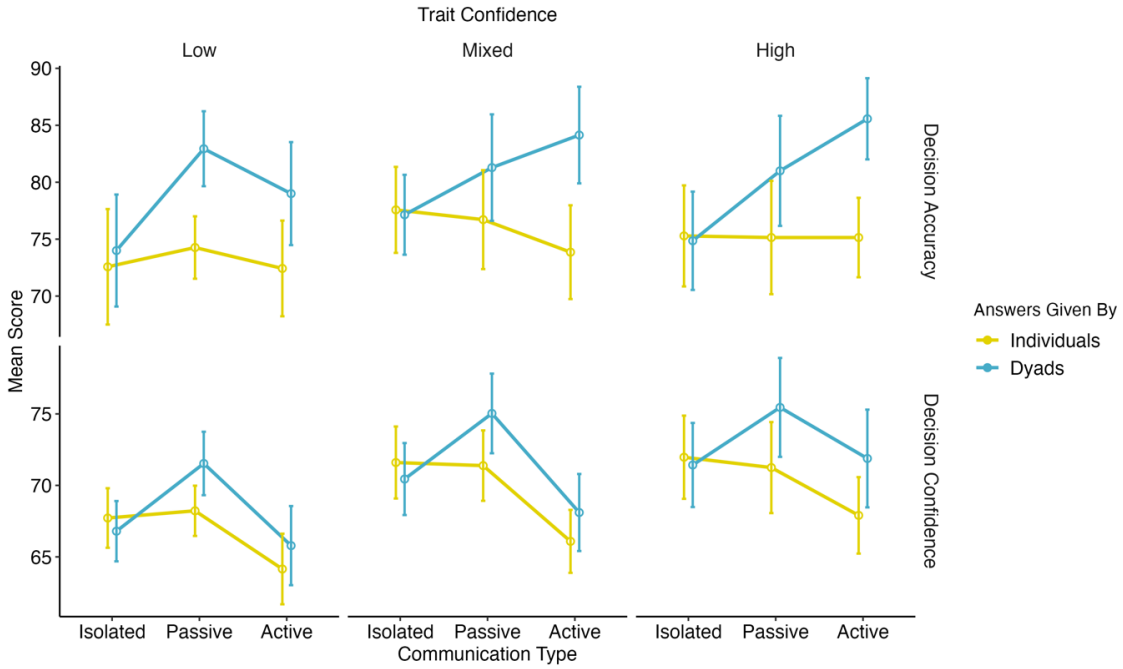


Table C7. *Random Effects for Communication Within Individuals and Dyads for Decision Accuracy and Decision Confidence*

Outcome	Individual			Dyad		
	Intercept SD	Passive SD	Active SD	Intercept SD	Passive SD	Active SD
Decision accuracy	13.77	14.98	11.95	6.75	6.01	9.35
Decision confidence	9.02	7.14	6.47	3.82	3.42	2.91

Table C8. *Baseline Comparisons for Individual Responses*

	Decision			
	Accuracy		Confidence	
Predictor	b	SE	*b*	SE
Low				
Passive vs Isolated	1.73	2.62	0.76	1.18
Active vs Isolated	-0.14	2.50	-3.57**	1.07
Active vs Passive	-1.87	2.58	-4.33***	1.13
Mixed				

Passive vs Isolated	-0.86	2.60	-0.22	1.16
Active vs Isolated	-3.71	2.50	-5.52***	1.07
Active vs Passive	-2.86	2.56	-5.30***	1.12
High				
Passive vs Isolated	-0.14	2.60	-0.72	1.16
Active vs Isolated	-0.14	2.50	-4.07***	1.07
Active vs Passive	0.00	2.56	-3.34**	1.12

*** $p < .001$, ** $p < .01$

Table C9. *Difference of the Differences for the Three-Way Interaction Effects*

			Decision			
			Accuracy		Confidence	
Trait Confidence	*Grouping*	*Communication*	b	SE	*b*	SE
Low	Dyad vs Ind	Passive vs Isolated	7.25**	2.25	4.23***	0.77
Low	Dyad vs Ind	Active vs Isolated	5.14*	2.24	2.56***	0.77
Low	Dyad vs Ind	Active vs Passive	-2.11	2.25	-1.68*	0.77
Mixed	Dyad vs Ind	Passive vs Isolated	5.00*	2.24	4.81***	0.77
Mixed	Dyad vs Ind	Active vs Isolated	10.71***	2.24	3.17***	0.77
Mixed	Dyad vs Ind	Active vs Passive	5.71*	2.24	-1.63*	0.77
High	Dyad vs Ind	Passive vs Isolated	6.29**	2.24	4.75***	0.77
High	Dyad vs Ind	Active vs Isolated	10.86***	2.24	4.52***	0.77
High	Dyad vs Ind	Active vs Passive	4.57*	2.24	-0.23	0.77
Communication	*Grouping*	*Trait Confidence*				
Isolated	Dyad vs Ind	Low vs Mixed	-1.86	2.24	-0.23	0.77
Isolated	Dyad vs Ind	Low vs High	-1.86	2.24	0.38	0.77
Isolated	Dyad vs Ind	Mixed vs High	0.00	2.24	0.61	0.77
Passive	Dyad vs Ind	Low vs Mixed	-4.11†	2.25	0.34	0.77
Passive	Dyad vs Ind	Low vs High	-2.82	2.25	0.90	0.77
Passive	Dyad vs Ind	Mixed vs High	1.29	2.24	0.56	0.77
Active	Dyad vs Ind	Low vs Mixed	3.71†	2.24	0.38	0.77
Active	Dyad vs Ind	Low vs High	3.86†	2.24	2.34**	0.77
Active	Dyad vs Ind	Mixed vs High	0.14	2.24	1.96*	0.77

*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$

Refitting the Final Models with Education Included as a Covariate

Table C10. *The Models for Decision Accuracy and Decision Confidence with Education as a Covariate*

		Decision			
		Accuracy		Confidence	
Predictor		b	SE	*b*	SE
Main Effects					
Grouping	Dyad vs Ind	5.22***	0.53	1.80***	0.18
Communication	Passive vs Isolated	3.33*	1.36	2.24**	0.64
	Active vs Isolated	3.12*	1.29	-2.67***	0.57
	Passive vs Active	-0.21	1.33	-4.91***	0.61
Trait Confidence	Mixed vs Low	2.88	2.16	3.33	1.69
	High vs Low	2.47	2.37	4.90*	1.85
	Mixed vs High	-0.41	2.21	1.58	1.74
EAT Acc		-0.07†	0.04	-0.06*	0.03
RAPM Acc		0.21***	0.05	0.15***	0.04
Education		-1.17	0.76	-0.63	0.59
Three-Way Simple Effects Interactions					
Trait Confidence	Communication				
Low	Isolated	1.43	1.58	-0.92†	0.54
Mixed	Isolated	-0.43	1.58	-1.16*	0.54
High	Isolated	-0.43	1.58	-0.54	0.54
Low	Passive	8.68***	1.60	3.31***	0.55
Mixed	Passive	4.57**	1.58	3.65***	0.54
High	Passive	5.86***	1.58	4.21***	0.54
Low	Active	6.57***	1.58	1.63**	0.54
Mixed	Active	10.29***	1.58	2.02***	0.54
High	Active	10.43***	1.58	3.98***	0.54

*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$

Selecting the Best Fitting Model for Testing the Emergence of Confidence Matching

Random Effect Structure

To select the best fitting model to report in the main text, we compared the same random effects structure as the models for decision accuracy and decision confidence with the addition of item number within each model. Each model had the same fixed effects, teammate's individual decision confidence, communication type (isolated vs passive vs active) and trait confidence (low-trait vs mixed-trait vs high-trait). We compared the models on AIC, BIC, and LRT. See Table C11 for the model comparisons.

Table C11. *Comparison of Models with Different Random Effects*

Model	AIC	BIC	LRT
1	46518	46659	-23238
2	46402	46557	-23178***
3	46468	46623	-23211
4	-	-	-
5	-	-	-
6	-	-	-
7	-	-	-
8	-	-	-

Model 1: no random effects
Model 2: (1 | individual) + (1 | item number)
Model 3: (1 | dyad) + (1 | item number)
Model 4: (1 | individual) + (1 | dyad) + (1 | item number)
Model 5: (1 + communication | individual) + (1 | item number)
Model 6: (1 + communication | dyad) + (1 | item number)
Model 7: (1 + communication | individual) + (1 | dyad) + (1 | item number)
Model 8: (1 + communication | individual) + (1 + communication | dyad) + (1 | item number)
*** $p < .001$, ** $p < .01$

Model 2, which included random intercepts for individuals and item number, provided the best fit based on AIC, BIC, and log-likelihood, with significant improvements over model 1. Model 3 did not improve the fit and models 4 through to 8 were singular and/or had convergence issues. Therefore, we retained model 2 as the final model.

Interaction effect structure

We compared 3 models to select the best fitting model for the main analyses. Each model had the same variables included as fixed effects: teammate's individual decision

confidence, communication type (isolated vs passive vs active), and trait confidence (low-trait vs mixed-trait vs high-trait). The models differed in their interaction effect structure. The models compared were: 1) main effects only (no interaction effects included); 2) the addition of all two-way interactions effects; and 3) the addition of three-way interaction effects. The results are presented in Table C12. Model 3 which included three-way interaction effects had the best fit, thus it was reported in the main text.

Table C12. *Comparison of Models with Different Interaction Effect Structures*

Model	AIC	BIC	LRT
1	46611	46686	-23295
2	46405	46533	-23183***
3	46402	46557	-23178*

Model 1: main effects only
Model 2: two-way interaction effects added
Model 3: three-way interaction effects added
*** $p < .001$, $p < .05$

Full Model Results for Testing the Emergence of Confidence Matching

The results for the final confidence matching model are presented in Tables C13, C14, and C15. Tests of each coefficient for all fixed effects are displayed in Table C13. Pairwise comparisons for main effects and simple effects contrasts for the interaction effects are shown in Table C14. Lastly, the difference of the differences contrasts for the three-way interactions are available in Table C15.

The model included random intercepts for both **dyad** and **item number**. The variance components indicated that **dyad-level differences** (**$\sigma^2 = 2.02$, SD = 1.42**) and **item-level differences** (**$\sigma^2 = 0.87$, SD = 0.94**) contributed **meaningful variation** in decision confidence changes. The **residual variance** was **$\sigma^2 = 94.90$, SD = 9.74**.

One difference-of-differences contrast reached significance, indicating that the difference in the effect of a teammate's individual decision confidence between passive and active communication type was significantly larger for low-trait confidence compared to

mixed-trait confidence (b = -0.11, SE = 0.04, t_{6169} = -2.83, p = .04). However, given the small effect size, the marginal significance level, and that none of the other contrasts approached significance, this result should be interpreted with caution.

Table C13. *Main and Interaction Effects for Confidence Matching Slopes*

Variable	Level	*b*	SE	*t*
Main Effects				
Teammate's Individual Confidence		0.15	0.01	20.30***
Communication	Isolated	-0.87	0.38	-2.29*
Communication	Passive	3.45	0.38	9.09***
Communication	Active	3.04	0.38	7.97***
Trait Confidence	Low	1.71	0.39	4.34***
Trait Confidence	Mixed	1.54	0.37	4.17***
Trait Confidence	High	2.38	0.40	5.92***
RAPM Acc	-	0.01	0.01	0.87
EAT Acc	-	-0.04	0.01	-2.48*
Three-Way Interactions Including Teammate's Individual Confidence				
Isolated	Low	0.02	0.02	1.14
Isolated	Mixed	-0.02	0.02	-0.86
Isolated	High	0.02	0.02	1.05
Passive	Low	0.28	0.02	13.46***
Passive	Mixed	0.20	0.02	10.85***
Passive	High	0.23	0.02	12.74***
Active	Low	0.19	0.02	8.54***
Active	Mixed	0.23	0.02	10.86***
Active	High	0.17	0.02	8.65***

*** $p < .001$, * $p < .05$

Table C14. *Simple Effects Contrasts for the Conditions for the Confidence Matching Model*

Variable	Contrast	*b*	SE	*t*
Main Effects				
Communication	Passive vs Isolated	4.32	0.49	8.74***
Communication	Active vs Isolated	3.91	0.50	7.89***
Communication	Active vs Passive	-0.41	0.50	-0.83
Trait Confidence	Mixed vs Low	-0.16	0.48	-0.34
Trait Confidence	High vs Low	0.67	0.55	1.21
Trait Confidence	High vs Mixed	0.83	0.50	1.65
Three-Way Simple Effects Interactions Including Teammate's Individual Confidence				
Trait Confidence: Low	Passive vs Isolated	0.25	0.03	8.63***
Trait Confidence: Low	Active vs Isolated	0.17	0.03	5.57***
Trait Confidence: Low	Active vs Passive	-0.08	0.03	-2.74†
Trait Confidence: Mixed	Passive vs Isolated	0.22	0.03	8.08***
Trait Confidence: Mixed	Active vs Isolated	0.25	0.03	8.65***
Trait Confidence: Mixed	Active vs Passive	0.03	0.03	1.11
Trait Confidence: High	Passive vs Isolated	0.21	0.03	8.24***
Trait Confidence: High	Active vs Isolated	0.15	0.03	5.71***
Trait Confidence: High	Active vs Passive	-0.06	0.03	-2.10
Communication: Isolated	Mixed vs Low	-0.04	0.03	-1.44
Communication: Isolated	High vs Low	-0.00	0.03	-0.16
Communication: Isolated	High vs Mixed	0.04	0.03	1.37
Communication: Passive	Mixed vs Low	-0.08	0.03	-2.87†
Communication: Passive	High vs Low	-0.05	0.03	-1.77
Communication: Passive	High vs Mixed	0.03	0.02	1.18
Communication: Active	Mixed vs Low	0.04	0.03	1.19
Communication: Active	High vs Low	-0.02	0.03	-0.67
Communication: Active	High vs Mixed	-0.06	0.03	-1.95

*** $p < .001$, † $p < .10$

Table C15. *Difference of the Differences Three-Way Interaction Effects for Confidence Matching*

Trait Confidence	Communication	*b*	SE	*t*
Mixed vs Low	Passive vs Isolated	0.03	0.04	0.89
Mixed vs Low	Active vs Isolated	-0.08	0.04	-1.90
Mixed vs Low	Active vs Passive	-0.11	0.04	-2.83*
High vs Low	Passive vs Isolated	0.04	0.04	1.11
High vs Low	Active vs Isolated	0.02	0.04	0.38
High vs Low	Active vs Passive	-0.03	0.04	-0.69
High vs Mixed	Passive vs Isolated	0.01	0.04	0.21
High vs Mixed	Active vs Isolated	0.09	0.04	2.41
High vs Mixed	Active vs Passive	0.09	0.04	2.30

* $p < .05$.

Selecting the Best Fitting Model for Confidence Matching Predicting the Change in Decision Accuracy

Random Effect Structure

We compared 6 models to select the best fitting model to include in the main analyses for the change in decision accuracy. Each model had the same fixed effects: confidence matching slope, communication type (isolated vs passive vs active), and trait confidence (low-trait vs mixed-trait vs high-trait). The models also included RAPM accuracy and EAT accuracy as covariates to control for differences in ability. The models differed in their random effects structure. See Table C16 for the model comparisons.

Table C16. *Comparison of Models with Different Random Effects*

	Decision accuracy Change		
Model	AIC	BIC	LRT
1	-758	-665	400
2	-756	-659	400
3	-	-	-
4	-	-	-
5	-	-	-
6	-	-	-

Model 1: no random effects
Model 2: (1 | individual)
Model 3: (1 | dyad)
Model 4: (1 | individual) + (1 | dyad)
Model 5: (1 + communication | dyad)
Model 6: (1 + communication | dyad) + (1 | individual)

Models 3-6 were singular as there was no variance in dyads and model 2 with random intercepts for individuals did not significantly improve the fit above a linear regression with fixed effects only. Inspection of the random effects showed very little variance across individuals ($\sigma^2 = 0.00$, SD = 0.01). Thus, model 1 was selected as the final model reported in the main results.

Interaction effect structure

We compared 3 models to select the best fitting model for the main analyses. Each model had the same variables included as fixed effects: teammate's individual decision confidence, communication type (isolated vs passive vs active), and trait confidence (low-trait vs mixed-trait vs high-trait). The models differed only in the complexity of the interaction structures. Specifically, the models included: 1) main effects only (no interaction effects included); 2) the addition of all two-way interactions effects; and 3) the addition of three-way interaction effects. The results are presented in Table C17.

Table C17. *Comparison of Models with Different Interaction Effect Structures*

	Decision accuracy Change		
Model	AIC	BIC	*F*
1	4912	4951	-
2	4913	4988	1.81
3	4915	5008	1.39

Model 1: main effects only
Model 2: two-way interaction effects added
Model 3: three-way interaction effects added

Model 2 which included two-way interaction effects showed a marginal improvement in model fit compared to model ($F_{8,600} = 1.81$, $p = .07$). Although, the improvement did not reach the significance threshold, the emergence of theoretically meaningful interaction effects suggests that including two-way interaction effects provides a more nuanced and accurate representation of the underlying processes involving confidence matching. In contrast, the inclusion of three-way interactions did not improve the fit. Thus, we reported model 2 in the main analyses. It offered a balance between parsimony and explanatory power. The two-way interaction model allowed us to examine meaningful patterns of moderation that were not detected in the main-effects-only model.

Full Model: Relationship Between Confidence Matching and the Change in Decision Accuracy

The results for the final confidence matching model are presented in Table C18 for the relationship between the change in decision accuracy and confidence matching at each level of the conditions including interaction effects. Pairwise comparisons for main effects and simple effects for the interaction effects are shown in Table C19.

Table C18. *Confidence Matching Slopes for Main and Interaction Effects*

Variable	Level	*b*	SE	*t*
Main Effects Slopes				
Confidence Matching		0.06	0.02	2.75**
Communication	Isolated	0.00	0.01	0.24
Communication	Passive	0.05	0.01	5.62***
Communication	Active	0.09	0.01	10.27***
Trait Confidence	Low	0.05	0.01	5.00***
Trait Confidence	Mixed	0.04	0.01	4.69***
Trait Confidence	High	0.06	0.01	5.49***
RAPM Accuracy	-	0.01	0.01	1.02
EAT Accuracy	-	-0.02	0.01	-4.52***
Slopes For Two-Way Interactions				
Confidence Matching	Isolated	0.01	0.05	0.13
Confidence Matching	Passive	0.15	0.04	3.74***
Confidence Matching	Active	0.04	0.02	1.69†
Confidence Matching	Low	0.09	0.03	2.73**
Confidence Matching	Mixed	0.07	0.04	1.84†
Confidence Matching	High	0.03	0.04	0.88

*** $p < .001$, ** $p < .01$, † $p < .10$

Table C19. *Contrasts for Full Confidence Matching Model for the Change in Decision Accuracy*

Variable	Contrast	*b*	SE	*t*
Main Effects				
Communication	Passive vs Isolated	0.05	0.01	3.54***
Communication	Active vs Isolated	0.09	0.01	6.42***
Communication	Active vs Passive	0.04	0.01	3.02**
Trait Confidence	Mixed vs Low	-0.01	0.01	-0.48
Trait Confidence	High vs Low	0.01	0.01	0.40
Trait Confidence	High vs Mixed	0.01	0.01	0.90
Two-Way Simple Effects Interactions				
Teammate's Individual Confidence	Active vs Isolated	0.14	0.07	2.10†
Teammate's Individual Confidence	Active vs Passive	0.03	0.06	0.58
Teammate's Individual Confidence	Mixed vs Low	-0.11	-0.05	-2.28†
Teammate's Individual Confidence	High vs Low	-0.02	-0.05	-0.53
Teammate's Individual Confidence	High vs Mixed	-0.06	-0.05	-1.33

Teammate's Individual Confidence	Active vs Isolated	-0.04	-0.05	-0.75

$^{***}p < .001$, $^{**}p < .01$, $^{\dagger}p < .10$

Summary of LPA Goodness of Fit Indices and Model Selection

LPA was performed for solutions with 2-6 classes (with 1 class as the default; see Table C20 and Figure C4 for a summary of the results) on the following 8 predictor variables: individual cognitive ability and trait confidence measured in the pre-screening study; and individual and dyadic accuracy and confidence measured on the GK test for each communication condition. Goodness of fit statistics were used to identify the number of latent profiles that best fit the data (Henson et al., 2007; Marsh et al., 2009). Assessment of the indices and examination of the profiles within each model suggested a 4-Class solution was the best model.

Figure C4. *A Summary of LPA Goodness of Fit Indices for 1-6 Classes*

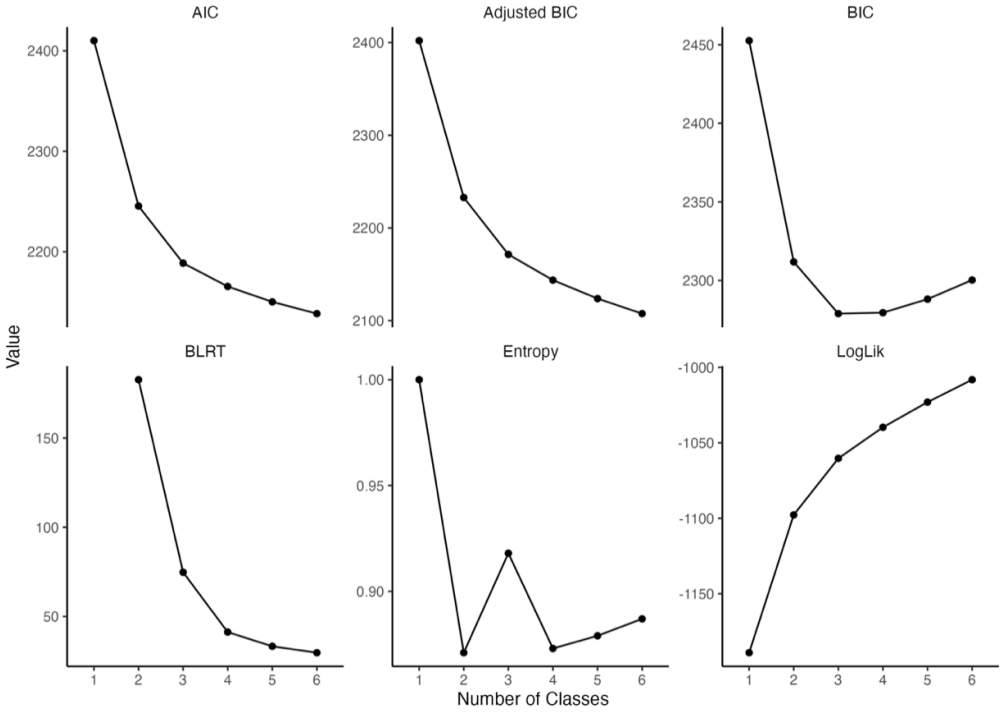


Table C20. *Goodness of Fit Statistics for All Latent Profile Analysis Models Tested (N = 105)*

Classes in the model	AIC	Adjusted BIC	BIC	Entropy	BLRT	LogLik diff		
1	2410	2402	2453	1.00	-	-91***		
2	2245	2233	2312	0.87	183*	-37***		
3	2189	2171	2279	0.92	75*	-21***		
4	2165	2144	2280	0.87	41*	-17***		
5	2150	2124	2288	0.88	33*	-15***		
6	2138	2108	2300	0.89	30*	-91***		
Class	Class counts and proportions for the latent classes		Average latent class probabilities for most likely latent class membership (row) by latent class (column)					
	Counts	Proportions	Class 1	Class 2				
Class 1	40	.38	.95	.05				
Class 2	65	.62	.03	.97				
	Counts	Proportions	Class 1	Class 2	Class 3			
Class 1	29	.18	.99	.01	.00			
Class 2	62	.59	.02	.97	.01			
Class 3	14	.13	.00	.05	.95			
	Counts	Proportions	Class 1	Class 2	Class 3	Class 4		
Class 1	37	.35	.96	.03	.00	.00		
Class 2	39	.37	.03	.91	.03	.03		
Class 3	17	.16	.04	.04	.92	.00		
Class 4	12	.11	.00	.05	.02	.93		
	Counts	Proportions	Class 1	Class 2	Class 3	Class 4	Class 5	
Class 1	27	.26	.96	.00	.04	.00	.00	
Class 2	37	.35	.01	.93	.05	.00	.01	
Class 3	22	.21	.04	.05	.90	.01	.00	
Class 4	11	.10	.00	.02	.01	.97	.00	
Class 5	8	.08	.00	.05	.02	.01	.92	
	Counts	Proportions	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	16	.15	.93	.00	.05	.02	.00	.00
Class 2	34	.32	.00	.92	.03	.02	.01	.01
Class 3	22	.21	.01	.04	.92	.02	.01	.00
Class 4	17	.16	.06	.04	.04	.86	.00	.00
Class 5	9	.09	.00	.03	.02	.00	.95	.00
Class 6	7	.07	.00	.02	.00	.00	.00	.98

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; BLRt = Bootstrap Likelihood Ratio Test; LogLik = Log Likelihood of the data, given the model. *p*-values of the chi-squared test between k and k-1 solutions.

* $p < .05$, *** $p < .001$

In particular, Akaike Information Criterion (AIC), sample adjusted Bayesian Information Criterion (BIC), and Bootstrap Likelihood Ratio Test (BLRT) declined from the 1-Class solution to the 6-Class solution but there tended to be a plateauing of model improvement around 3- to 4-Classes. All solutions had acceptable entropy values (>.80 is considered good, Clark & Muthén, 2009) and entropy was greatest for the 3-class solutions (.92). Entropy was reverse coded so 1 = complete certainty of classification and 0 = complete uncertainty. The Log-likelihood of the data, given the model (LogLik) decreased for each model - indicating improved the fit for each subsequent solution – and chi-square tests demonstrated that the difference between models was significant. Evaluation of the goodness of fit indices indicated that all solutions were viable candidates for selection with a slight advantage for the 3- and 4-class solutions which appear to offer a balance between model complexity and incremental gains on the goodness of fit indices.

Next, we compared the solutions on the proportion of class membership and the estimated accuracy of classification. The 2-, 3-, and 4-class solutions had an adequate proportion of members (>.10) in each profile, but the 5- and 6-class solutions were below this membership threshold. Furthermore, the estimated accuracy of classification was examined using the average latent class probabilities for the most likely class membership. For the 2- to 5-class solutions all probabilities were 90% or higher, which is acceptable, however, one class was below this threshold for the 6-class solution. See Table C20: high values on the diagonal and low values off the diagonal indicate goodness of classification).

To select a model from the 2-, 3-, and 4-class solutions, we examined the interpretations of the profiles within each model (see Figure C5). All models included high (class 1) and moderate (class 2) performance dyads. The model with 3-classes included an additional profile that was significantly more accurate for active communication (M = -0.73, SD = 2.15, CI = [-1.14, -0.32]) than passive communication (M = -1.41, SD = 3.21, CI = [-

2.03, -0.80]). The model with 4-classes included another profile where the opposite was true: accuracy was significantly higher for passive communication (M = 0.11, SD = 4.22, CI = [-0.69, 0.92]) than active communication (M = -1.81, SD = 6.56, CI = [-3.07, -0.56]. These two additional profiles identified participants who benefitted from different types of communication and provided meaningful insights into systematic outcomes for dyads. Thus, we selected the 4-class solution as the best fitting model because it provided important information beyond the 2- and 3-class solutions. All subsequent analyses were based on this model.

Figure C5. *Comparison of the 2-, 3-, and 4-Class LPA Solutions*