ARL-TR-9677 • Apr 2023

DEVCOM ARMY RESEARCH LABORATORY

Measuring Cohesion in Human-Autonomy Teams: Development of a Cohesion Scale for Use in Human-Autonomy Team Research

by Catherine Neubauer, Shan Lakhmani, Andrea Krausman, Sean M Fitzhugh, Samantha Berg, Ericka Rovira, Jordan Blackman, Trinity Garay, and Daniel Forster

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9677 ● APR 2023

Measuring Cohesion in Human-Autonomy Teams: Development of a Cohesion Scale for Use in Human-Autonomy Team Research

Catherine Neubauer, Shan Lakhmani, Andrea Krausman, Sean M Fitzhugh, and Daniel Forster
DEVCOM Army Research Laboratory

Samantha Berg
Oakridge Associated Universities

Ericka Rovira, Jordan Blackman, and Trinity Garay
US Military Academy at West Point

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
April 2023	Technical Report	10/1/2020	9/30/2022

4. TITLE AND SUBTITLE
Measuring Cohesion in Human-Autonomy Teams: Development of a Cohesion Scale for Use in Human-Autonomy Team Research

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Catherine Neubauer, Shan Lakhmani, Andrea Krausman, Sean M Fitzhugh, Samantha Berg, Ericka Rovira, Jordan Blackman, Trinity Garay, and Daniel Forster

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FA Aberdeen Proving Ground, MD 21005	ARL-TR-9677

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs: Catherine Neubauer, 0000-0002-6686-3576; Shan Lakhmani, 0000-0001-6052-439X; Andrea S Krausman, 0000-0003-1955-8867; Sean M Fitzhugh, 0000-0002-6283-2895; Samantha Berg, 0000-0001-5774-8747; Ericka Rovira, 0000-0002-4820-5828; Daniel Forster, 0000-0001-8351-2009

14. ABSTRACT
Cohesion is a critical aspect of human-autonomy team function and effectiveness, yet there is a need for more robust methods to adequately measure this construct. This report documents the process used to develop and validate a new team cohesion scale, specifically directed toward use within human-autonomy teams. Here we describe the phases of the development process including item development, scale item evaluation, content validation, and an online scale validation study to further reduce the number of items. Taken together, the results of these analyses highlight several items with very good measurement properties, especially from the items designed to assess perceived complementarity, morale, leadership direction, and perceived efficacy. However, some items with excellent properties still belong to measurement scales with apparent issues. Overall, these results will help guide recommendations for future measures of cohesion in human-autonomy teams.

15. SUBJECT TERMS
cohesion, human–autonomy teams, group dynamics', human–autonomy interaction, autonomous teammates, Humans in Complex Systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	79

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Catherine Neubauer	(954) 258-2287

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Acknowledgments

The authors would like to thank the following subject matter experts and reviewers for their helpful feedback: Nancy Cooke (Arizona State University), Joseph Coyne (Naval Research Laboratory), Arwen DeCostanza (DEVCOM Army Research Laboratory), Gregory Funke (Air Force Research Laboratory), Joseph Lyons (Air Force Research Laboratory), Julie Marble (Johns Hopkins University), Gerry Matthews (George Mason University), Nathan McNeese (Clemson University), Matthias Scheutz (Tufts University), Charlene Stokes (DEVCOM Army Research Laboratory), Katia Sycara (Carnegie Mellon University), and Matthias Scheutz (Tufts University).

1. Introduction

Since it has been posited that the foundation of human-autonomy team (HAT) literature is built on that of interpersonal, or human teams (Morrow and Fiore 2012), we hereby extend the interpersonal definition of teams, "two or more [team members that] interact dynamically, interdependently, and adaptively toward a common and valued goal, objective, or mission, ... have been each assigned specific roles or functions to perform, and have a limited span of membership," to HATs as well (Salas et al. 1992, p. 4). For the scope of this work, a HAT consists of one or more human teammates, coupled with one or more autonomous systems, or intelligent agents (IAs), that collaborate to accomplish a task or goal (Demir et al. 2019).

To build on this new team dynamic, recent advances in artificial intelligence have endowed autonomous systems and other IAs with a greater capability for independence as well as interdependence, thus moving technology away from roles that simply support or augment human performance in limited ways, to the adoption of roles as actual team members that truly extend the overall team dynamics and capabilities (Phillips et al. 2011; Demir et al. 2019). In this context, IAs are autonomous entities, with the ability to observe and act on their environment, as well as conduct activities toward achieving both individual and collective goals (Russell and Norvig 2010). IAs can be computer-based entities (i.e., embedded agents) or physical entities (i.e., embodied agents, a.k.a. robots). Within both embedded and embodied agent systems, there can be multiple tasks performed using multiple types and levels of autonomy, making it difficult for human teammates to understand the agent's actions or decision-making processes. As such, successful integration of humans and autonomous systems as team members requires each to understand the other's reasoning, actions, and intentions (Chen et al. 2018; Schaefer et al. 2017). This shared understanding is foundational for teamwork and the development of critical team states such as trust and cohesion.

1.1 Team Cohesion Definition and Gaps in Human-Autonomy Teams

Team cohesion has been described as the most important determinant of team success (Carron and Brawley 2000). Research suggests that team cohesion benefits team productivity by increasing performance and has a positive psychological impact on team members as well (Beal et al. 2003; Mathieu et al. 2015; Neubauer et al. 2016). Although thoroughly researched in human teams, team cohesion has not yet been explored in HATs, even though the last decade has burgeoned with interest in teaming humans and robotic or autonomous systems, especially in the

military (Barnes and Evans 2010). Compared to human-human teams, communication, organizational hierarchy, and collaboration work differently in HATs (Lakhmani et al. 2022). This could potentially pose a challenge for working together effectively as well as for developing critical team processes. For these reasons, it is necessary to understand what aspects of team cohesion relate to HATs and how they can be measured.

The operational definition of cohesion is "the shared bond/attraction that drives team members to stay together and to want to work together" (Salas et al. 2015). Understanding cohesion also requires us to "understand the levels of trust that will enhance usage and effective human robot interaction" (Schaefer 2016, p. 216). As a construct, it is most efficient to separate cohesion into factors and subdimensions due to its expansiveness (Griffith 1988; Zaccaro 1991; Griffith and Vaitkus 1999; Dion 2000; Salas et al. 2005). Because of this, we completed a thorough literature review on human team cohesion and divided the construct of cohesion into several dimensions and subdimensions that have been commonly accepted as factors of cohesion in the literature (Lakhmani et al. 2022). These factors served as the foundation for a subjective scale specifically designed to measure the unique characteristics of cohesion in HATs.

In HATs, aspects of cohesion should be treated and defined differently than in human-human teams because social interactions between humans and autonomous agents are unique and play an important role in team coordination (Walliser 2019). As such, HATs must maintain adequate levels of team cohesion to best allow the team to perform well in terms of mission success and maintain psychological well-being. Team cohesion and solidarity are vital factors that give teams the ability to perform better (Mudrack 1989; Beal et al. 2003; Chiocchio and Essiembre 2009). These factors are important to consider as we seek to "team" humans and autonomous systems. As technology pervades every facet of our lives, we must learn how to leverage its benefits while minimizing its weaknesses. Within this line of research, it is also crucial to expand upon previously developed metrics of cohesion to determine if previously used measures of cohesion for human-human teams are still appropriate or if they need to be adapted to better fit the evolving dynamic within HATs.

With respect to cohesion measures, although there are existing methods for assessing team cohesion, there are currently no self-report scales that specifically target HATs. Thus, the goal of the current effort was to develop a new subjective cohesion scale that will allow us to evaluate team cohesion, solidarity, and other factors that contribute to HAT effectiveness. This report documents the scale development process that followed a three-phase approach outlined in Boateng et al. (2018). In Phase 1 item development, scale items were generated, and the

content was validated by subject matter experts (SMEs). In Phase 2, the scale was created based on SME feedback, and in Phase 3, an online study was conducted to evaluate the reliability and validity of the scale. The items we evaluated allowed us to determine which factors played the biggest role in maintaining high-functioning teams.

1.2 Scale Development Process

The scale development process can be broken down into three phases: 1) item development, 2) scale development, and 3) scale evaluation (see Fig. 1).

1.2.1 Phase 1: Item Development

Item development includes two parts: ***initial item pool generation*** and ***content validation***.

Item generation is the way researchers should build theoretical support for the initial item pool (Morgado et al. 2018). The two methods used are deductive and inductive. The deductive method includes item generation based off an extensive literature review and pre-existing scales (Morgado et al. 2018). The inductive method includes item development on qualitative information regarding a construct obtained from opinions gathered from an expert target population (Morgado et al. 2018). For this work, the initial item pool developed was drawn from existing human cohesion scales and resulted in 134 items, which were used for the content validation experiment.

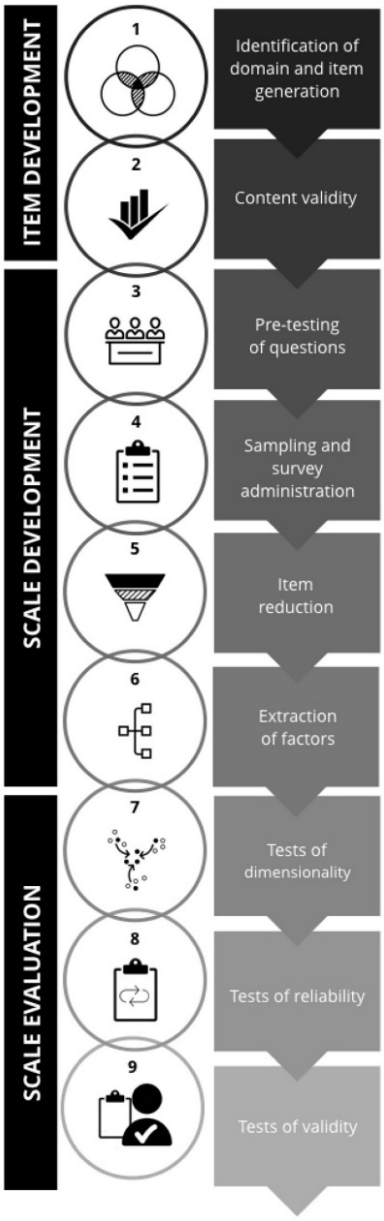

Fig. 1 An overview of the three phases and nine steps of scale development and validation (Boateng et al. 2018)

A comprehensive literature review was conducted to understand the construct of team cohesion more accurately. Following this effort, it was found that some approaches suggest that cohesion is composed of two dimensions: the direction of cohesion—vertical (superior-subordinate relationships) or horizontal (peer relationships); and the functions of cohesion—instrumental (task-based) or affective-based (relating to interpersonal support) (Griffith 1988; Dion 2000). Another approach comes from military cohesion, which divides cohesion into four related components, composed of **primary** (e.g., vertical and horizontal cohesion)

and **secondary** (e.g., organizational and societal cohesion) hierarchical components (Siebold 2006). Iterating upon a previous aggregation of these models of cohesion (Lakhmani et al. 2022), we designed a cohesion framework to organize our item pool (see Fig. 2).

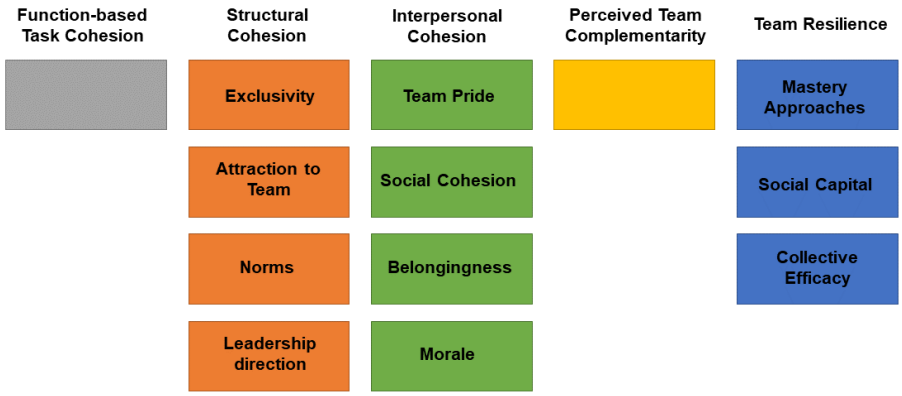

Fig. 2 Visualization of 5-D model of cohesion used in the current HAT Cohesion Scale. Underneath each dimension of cohesion are listed any subdimensions.

Following these multidimensional representations, our scale has been designed to include the following five factors of cohesion: Function-based Cohesion, Structural Cohesion, Interpersonal Cohesion, Perceived Team Complementarity, and Team Resilience. Descriptions for these factors and associated subfactors or dimensions are as follows:

1) Function-based Task Cohesion

   Instrumental or task cohesion is action-oriented or proactive and has been defined as a group's shared commitment or attraction to the group task or goal or the group's capacity for teamwork (Siebold 1999). Further, it is a shared understanding of and commitment to group tasks and goals (Beal et al. 2003). This is perhaps the most straightforwardly applicable to teaming with intelligent agents, as this type of teaming requires a level of joint mission parameters and goals.

2) Structural Cohesion (four subdimensions)

   a. Exclusivity

      Exclusivity, according to self-categorization theory, reflects the extent to which individuals adhere (via attitudes, behaviors) to group norms that characterize the in-group and distinguish themselves from out-groups (Hogg 1992). Members of different groups who perceive themselves as belonging to a superordinate group will

increasingly recognize previous outgroup members as part of a larger, more inclusive group (Gaertner and Dovidio 2009). When subgroup differences are not acknowledged within the superordinate group, however, subgroup members will grow more defensive of their own group at the expense of other groups (Crisp et al. 2006).

b. Attraction to Team/Resistance to Leaving

Interpersonal attraction is a shared liking for or attachment to the members of a group; it is important to note, however, that liking group members is not the same thing as liking the group (Beale et al. 2003; Abrams and Rosenthal-von der putte 2020). Rather, liking the group is more closely associated with another component of cohesion: group pride. Interpersonal attraction has been described as central to the cohesiveness of small groups, enough so that some unidimensional conceptualizations of cohesion equated the two (Dion 2000; Lott and Lott 1965). While this approach has fallen out of fashion, interpersonal attraction has been shown to have a meaningful correlation with performance (Beal et al. 2003).

c. Norms

Norms are standards for evaluating what behaviors are and are not acceptable within a group to establish expectations of team members (Forsyth 1999). Norms are complementary to task cohesion in that norms can be used to focus the effort team members put toward group tasks (Carron and Spink 1993). Because norms can be used to enhance (or degrade) performance, they serve as moderators in the relationship between cohesion and group performance (Carron and Spink 1993; Langfred 1998).

d. Leadership Direction of Cohesion

The primary dimension of cohesion, that is, the direction of cohesion, emphasizes the role of hierarchy in team cohesion (Griffith 1988). For the current effort, this dimension contrasts superior-subordinate relations (Siebold and Kelly 1988; Dion 2000). This distinction is often included in examinations of cohesion in the military (Siebold and Kelly 1988; Grossman 2014).

3) Interpersonal Cohesion (four subdimensions)

a. Team Pride

Group pride, or the support for group beliefs and group representation, seems to be a heavily affect-based component of cohesion. It is the extent to which group members exhibit liking for the status or the ideologies that the group supports or represents (Beal et al. 2003). It has also been defined as the shared importance of being a member of the group and has a long-standing importance within the cohesion literature (Mullen and Copper 1994). In the cohesion-performance relationship, group members will work harder for the pleasure of belonging to a high status, successful group, yet it is not a predictor in the cohesion-performance relationship (Mullen and Copper 1994). As a factor, group pride does not seem to have the prominence in the cohesion literature that it once did; therefore, we don't expect it to be particularly relevant to HAT cohesion, especially as autonomy gets more common (and this becomes less of a potential status symbol).

b. Social Cohesion

Social cohesion is considered the other major function of cohesion (Griffith and Vaitkus 1999; Dion 2000). Social cohesion, also described as interpersonal cohesion, is the group members' attraction to or liking of the group and their trust in group members (Evans and Jarvis 1980; Siebold 1999; Craig and Kelly 1999). Individual-level indicators of social cohesion include the following:

> (a) individuals' membership attitudes (their desire or intention to remain in a group, their identification with or loyalty to a group, and other attitudes about the group or its members); and (b) individuals' membership behaviors: their decisions to sever, weaken, maintain, or strengthen their membership or participation in a group, their susceptibilities to interpersonal influence, and other behavioral indicators of commitment and attachment to the group (Friedkin 2004: 410).

Social cohesion is considered an integral aspect of well-functioning groups (Ahronson and Cameron 2007). However, researchers have described other components of cohesion, outside of the functional or directional dimensions, whose elements can be grouped under

social cohesion, such as belongingness or morale (Dion 2000; Grossman 2014).

c. Belongingness

Stemming from the work of Bollen and Hoyle (1990), *belongingness* is the degree to which members of a group are attracted to each other (Salas et al. 2015). It is grounded in both group members' cognitive appraisals of the degree to which they belong in a group and their affective responses to such appraisals (Bollen and Hoyle 1990; Grossman 2014). This aspect of cohesion has been considered fundamental to the existence of a group, such that a sense of belonging to a group is a prerequisite to any other group characteristic (Bollen and Hoyle 1990). Research into belongingness has shown that it correlates with social outcomes and social self-esteem (Dion 2000).

d. Morale

Morale, along with belongingness, stems from Bollen and Hoyle's (1990) work with perceived cohesion. Morale refers to the global affective response, positive and negative, associated with being in a group (Bollen and Hoyle 1990). It can also be defined as an individual's degree of loyalty to fellow group members and their willingness to endure frustration for the group (Salas et al. 2015). This factor has a temporal component as well, as it shapes group member responses to conflicts or setbacks (Dion 2000; Grossman 2014). While this factor is highly correlated with belongingness, it is in fact distinct; one example that illustrates this distinction is that of a natural disaster hitting a city, which may increase one's feelings of belongingness to that city, while simultaneously reducing morale (Bollen and Hoyle 1990).

4) Perceived Team Complementarity

Complementarity refers to the diversity of skill sets that group members bring to the larger team and how these skill sets can meet the needs of the environment (Muchinsky and Monahan 1987). Complementarity, a recently postulated dimension of cohesion (Lakhmani et al. 2022), is composed of some social and task cohesion (e.g., a robotic/autonomous system must have skills that complement/augment their team's skills or abilities to complete a required task). It is the assumption that teams become cohesive

when members exhibit different, but complementary skills, allowing some team members to make up for other team members' weaknesses.

5) Team Resilience (three subdimensions)

> *Resilience* is fundamental to team cohesion and subsequent success when teams encounter environmental and team stressors (Berg et al. 2021). Team resilience is defined as "a multi-phasic process in which members of the unit deliberately and collectively apply skills, abilities, and resources to prepare the unit for adversity by planning and anticipating adverse events, successfully respond to challenging events by withstanding or adapting to stressors, and recover after the event, which involves the unit returning to homeostasis (e.g., bouncing back) or an improved state through post-event learning and growth" (Cato et al. 2018, p. 53).

Additionally, it can be argued that resilience is a key feature for the development of highly cohesive and trusted human teams (Gittell et al. 2006; Norris et al. 2008). In fact, resilience is sometimes viewed as a combination of team states including collective efficacy, shared mental models, and familiarity (Bowers et al. 2017). This area of work is particularly relevant for teams in extreme environments where cohesion is impacted differently than in teams under normal conditions (Salas et al. 2017). For example, individuals working in extreme environments tend to exaggerate issues, which may lead to group impairment when increased tension and perception of team problems negatively impact team cohesion (Stuster 1996). However, military unit cohesion has been shown to counteract these extreme environment stressors (Williams et al. 2016). In recent years, there has been a push to integrate robotic systems as team members in military operations to increase efficiency and decrease risk to Warfighters (Barnes and Evans 2010). These HATs are especially effective for open-ended and complex conditions where aspects of a task are not always mapped or planned out (e.g., combat situations; Chen and Barnes 2014), by aiding in information planning, task planning and allocation, and team operations (Sycara and Sukthankar 2006). However, it is paramount to understand how the incorporation of robotic systems to human teams may disrupt the teams' homogeneity and subsequent cohesion and resilience (O'Reilly III et al. 1989; Smith et al. 1994).

There are **three subdimensions** within the theoretical framework of **team resilience** (Morgan et al. 2013) including:

a. Mastery Approaches

"The shared attitudes and behaviors of the team members that promote an emphasis on team improvement" (Morgan et al. 2013, p. 553). The subdimensions of mastery approaches are **team learning** and **team flexibility**.

i. Team Learning Orientation:

"Activities carried out by team members through which a team obtains and processes data that allow it to adapt and improve" (Edmondson 1999, p. 351).

ii. Team Flexibility Orientation:

Team members' ability to assess and adjust their behavior and structure with the goal of functioning effectively in stressful situations (Griffin 1997).

b. Social Capital

"Features of social life-networks, norms, and trust, which enable participants to act together more effectively to pursue shared objectives" (Putnam 1995, p. 56). Includes the subdimension of shared language.

i. Shared Language:

Reflects the team culture and how it impacts the formation of social relationships.

c. Collective Efficacy

"Group's shared belief in its ability to organize and execute the actions required to reach certain levels of achievement" (Bandura 1997, p. 477). Described in the subdimension of **Perceived Efficacy for Collective Team Action**.

i. Perceived Efficacy for Collective Team Action:

Reflects the team's perceived ability to complete actions as a unit and face adversity collectively.

1.2.2 Initial Item Pool

For the current scale development, a thorough literature review was conducted on existing team cohesion scales to adapt those items that best fit this framework (Berg et al. 2021). Our result was an initial pool of 134 items, containing the following dimensions: function-based task cohesion, structural cohesion (Griffith 1988), interpersonal cohesion (Salas et al. 2015), perceived team complementarity (Piasentin and Chapman 2007), and team resilience (Cato et al. 2018). Many of these dimensions are well known within the human-teaming cohesion literature; however, for the current scale development effort, the addition of two further factors was considered (i.e., items relating to the dimensions of complementarity and resilience). In this context, it has been argued that complementarity occurs when "an individual possesses unique characteristics that are perceived to be different from others' characteristics, yet [are] valuable to the organization" (Piasentin and Chapman 2007, p. 234). For perceived complementarity, we adapted 18 items from Oosterhof et al. (2009) and Piasentin and Chapman (2007). For the subdimension relating to team resilience, we adapted 20 scale items from Sharma and Sharma (2016) (see Berg et al. 2021 for further reading on item adaptation).

Content Validation. The second step of item development includes *theoretical analysis*. In this step, content validity assessment is required because inferences are made based on final scale items (Morgado et al. 2018). This assessment included the opinions of SMEs or the user population.

Our initial item pool was sent to 11 SMEs from academic and government settings who are known for researching team cohesion and/or HATs. These SMEs completed content validation procedures by rating each of the 134 items using a 3-point ordinal scale (0 = "should not be included in the scale"; 1 = "important to include in the scale"; 2 = "extremely important to include in the scale"). Additionally, SMEs provided qualitative, written feedback and recommendations for items. The items were analyzed using the Content Validity Ratio and procedures outlined by Lawshe (1975). The formula for item-level SME agreement yields values ranging from +1 to –1; positive values indicate that at least half of the SMEs rated the item as "Extremely Important." With 11 SMEs, the threshold value for item removal was determined to be .59, to ensure that the SME agreement is unlikely to be due to chance, which resulted in an item reduction from 134 to 82 items (Appendix A) that were evaluated in the online study for Phase 2.

1.2.3 Phase 2: Scale Development

The goal of Phase 2 is to reduce the item pool and identify potential factors within the scale. There are four steps in this process: ***pre-testing***, scale ***administration and sample size***, ***item reduction analysis***, and ***extraction of factors*** (described in detail in Boateng et al. 2018). ***Pre-testing*** ensures that the items are meaningful to the target population, by eliminating unrelated or poorly worded items and revising remaining items so that they will be easily understood by the target population. Pre-testing, for this scale, occurred during the previously mentioned SME review. Part of the qualitative feedback included revisions to the original 134 scale items so that they would better adhere to the context of HAT cohesion.

Once the SME feedback was implemented, we initiated data collection, taking the *survey administration* and *establishing adequate sample size* steps. This process is described in Section 2. This data collection also serves as the initial baseline for a test-retest setting, which is needed for effective reliability testing during the evaluation stage.

Once the data was collected, we began the *item reduction analysis* step, where we used a combination of confirmatory and exploratory factor analysis methods. First, we used confirmatory analysis to test whether the items designed to measure distinct constructs fit under a series of single-factor models. If the model did not fit well, we then removed items until we obtained a well-fitting model, then tested the model for longitudinal invariance. When testing for invariance, we could examine whether item responses have the same factor structure across contexts, and we could identify whether specific items were not invariant and should be removed. This process is discussed in more detail in Section 3. Finally, we used exploratory factor analysis on the reduced scale to test whether the novel items loaded on our criterion measure of team cohesion—if items clustered with the established scales, they were considered for removal, whereas if items clustered onto a distinct factor, they were retained to measure that novel subdimension.

1.2.4 Phase 3: Scale Evaluation

The final phase is psychometric analysis for scale evaluation. This analysis assesses whether the scale has construct validity (what the instrument is measuring) and reliability (score consistency) (Morgado et al. 2018). The next section will outline the validation study, which was employed to allow for the scale evaluation process.

2. Method

The overall objective of this line of research was to create an instrument with the capability to measure and calibrate team cohesion within a HAT. The following

section will outline Phase 2 of the scale development process by describing the current experiment to determine if our new cohesion scale is valid and reliable as well as reducing the item pool so it can be used in further HAT research studies.

2.1 Instrumentation and Facilities

The current online study was hosted on the Department of Behavioral Sciences and Leadership's (US Military Academy [USMA]) Qualtrics account. All data were collected and stored on the servers associated with the BS&L Qualtrics account. Participants logged into Sona using USMA provided laptops. Sona provided a description of the research and allowed participants to easily sign up for research studies. Sona provided participants a link to Qualtrics, allowing participants access to the vignettes, video clips, and survey.

2.2 Materials, Tests, Task, and Stimuli

The following questionnaires were completed by participants prior to and during the experimental task:

2.2.1 Questionnaires

- *Demographics Scale:* This brief questionnaire was created by the researchers to collect relevant non-personally identifiable demographic information as well as vision and hearing status. In addition, participants were asked if they had experience with gaming equipment and if they experience motion sickness with non-immersive displays. This was given once at the beginning of the experiment.
- *Game Experience Measure*: Experience with and knowledge of video games has been theorized to influence future gaming performance (Taylor et al. 2009). This coupled with the strong association between cohesion and team play, both virtual and physical, suggests that game experience should be measured (Salas et al. 2015). The questions in this self-report measure (Taylor and Barnett 2011) assessed the participants' general video game experience and were given once at the beginning of the experiment.
- *Mini-IPIP Personality Inventory:* The Mini-IPIP scale (Donnellan et al. 2006) personality assessment is used to measure the Big Five personality traits: Agreeableness, Extraversion, Intellect, Conscientiousness, and Neuroticism. It is a 20-item short form of the 50-item International Personality Item Pool – Five Factor Model (Goldberg 1999). Limited works in HRI suggest that personality traits are highly correlated with trust (Looije et al. 2010) and thus may also be indicative of team cohesion emergence.

Donnellan et al. (2006) found consistent and acceptable internal consistencies across five studies, with Cronbach's alpha coefficients at or above 0.60. The Mini-IPIP was administered once at the beginning of the experiment.

- *System Trustworthiness*: Participants rated 7-point Likert-type questions, rating the perceived level of intelligence of the robot, perceived level of automation, perceived trustworthiness, perceived safety, and perceived use/teaming. These items are adapted from Schaefer et al. (2012), used in prior Wingman studies (ARL-18-165; Schaefer et al. 2019) and was given once, prior to the start of the experimental task.
- *Negative Attitudes toward Robots Scale (NARS)*: The 7-point NARS (Nomura et al. 2006) measures negative attitudes toward robots based on emotions in interactions, social influence, and situational influence. It was administered once, before the experimental task began. It has been previously used across multiple domains of HRI and has been shown to predict interaction and explain individual differences in participants' behavior. For a review of studies using the NARS, see Tsui et al. (2010).
- *Immersive Tendencies Questionnaire (ITQ):* The Immersive Tendency Questionnaire (Witmer and Singer 1998) is a commonly used scale in the presence literature. This measure is used to identify individuals' levels of immersion and consists of 32 items, which break down into subscales that measure tendencies regarding involvement, focus, and gaming. High scores indicate more immersion and lower scores indicate less immersion. The reported reliability is $\alpha = 0.81$ (Witmer and Singer 1998). This was given once at the beginning of the experiment.
- *The Positive and Negative Affect Schedule Extended (PANAS-X)*: The PANAS-X measures both Positive Affect (PA) and Negative Affect (NA), as well as 11 primary affects labeled: Fear, Sadness, Guilt, Hostility, Shyness, Fatigue, Surprise, Joviality, Self-Assurance, Attentiveness, and Serenity. The PANAS-X includes eight different temporal instructions ranging from the following: "Right Now"; "Today"; "Past Few Days"; "Past Week"; "Past Few Weeks"; "Past Month"; "Past Year"; and "In General" (Watson and Clark 1999). As a measure of transitory emotions, the instructions ask respondents to rate how they feel "Right now (at the present moment)" scored on a 5-point Likert-type intensity scale ranging from "Very slightly or not at all"; "A little"; "Moderately"; "Quite a bit"; to "Extremely." Extending their assessments to longer-lasting mood states, respondents are next instructed to indicate to what extent they have felt this

way “during the past few weeks.” With instructions to respond as to how they felt “during the past year” the PANAS-X is measuring dispositional affect dimensions. Given this wide range of temporal instructions, the PANAS-X can provide greater flexibility in the measurement of affects, as compared with state trait measures. This was given once, before the experimental task began.

- *The Group Cohesion Questionnaire* (Carless and De Paola 2000) is an adaptation of the 18-item Group Environment Scale (GEQ) originally developed for use with sports teams. The scale consists of 10 items measuring task cohesion, social cohesion, and interpersonal attraction using a 9-point Likert-type scale ranging from strongly disagree (1), to strongly agree (9). Wording for some of the original GEQ items was changed to reflect an organizational setting with references to work teams rather than sports teams and is henceforth referred to as the GEQ-10 in this report. This was administered after each video clip and served as our criterion measure of cohesion, against which our novel items from each subscale were evaluated.

2.2.2 Experimental Task

Due to COVID-19, in-person data collection was not possible. Therefore, the current study was conducted online, through Qualtrics. The task primarily consisted of written vignettes describing HATs who work together. Additionally, participants viewed short video clips of the same nature. Specifically, these clips featured high and low cohesive teams consisting of human and robot team members performing a collaborative task that was specific to each of the subfactors of cohesion described previously.

Using a laptop, participants viewed short video clips of low and high cohesive HATs from the television show *Star Wars: The Clone Wars*. Because actual HATs do not widely exist, we decided to take a vignette approach. Several members of the research team independently identified levels and types of cohesion in 40 clips from the show before collectively narrowing them down to 18 clips based on their clarity and representation of the specific subdimensions of cohesion. As such, these video clips are reflective of the different subdimensions of cohesion outlined earlier in our team cohesion scale. Following video clip presentation, participants were asked to rate their perceived level of the team’s cohesion using our scale, which was rated on a 7-point Likert scale. This consisted of evaluating the team via the items in our item pool, in addition to rating the items from the shortened versions of the GEQ-10.

2.3 Subjects and Sample Size

2.3.1 Subjects

We collected data from 294 USMA cadets. All cadets had completed basic military training; some cadets are former enlisted Soldiers. At the time of data collection, cadets were currently enrolled at USMA in the Introduction to Psychology for Leaders Course (PL300). Cadets ranged in age from 18 to 23 years (M = 19.97, SD = 1.49), and represented all 50 states. Of the sample, 24.5% identified as female, 71.4% identified as male, and the remaining 4.1% did not choose to reveal this information.

Cadets were recruited using the USMA Sona system and were able to earn extra credit as part of their participation in research conducted by cadets and faculty (i.e., 10 points = 1 h). If cadets did not wish to participate in this study, they could choose to complete reviews of journal articles (1 article = 10 points).

2.3.2 Sample Size Justification

Sample size justification for scale development varies widely from a minimum of 2 to a maximum of 10 participants per scale item (Anthoine et al. 2014). Others have suggested sample sizes that are independent of the number of survey items, recommending a range of 200–300 as appropriate for factor analysis (Guadagnoli and Velicer 1988; Comrey 1988).

According to Boateng et al. (2018),

> there is no single item-ratio that works for all survey development scenarios. A larger sample size or respondent: item ratio is always better, since a larger sample size implies lower measurement errors and more stable factor loadings, replicable factors, and generalizable results to the true population structure (MacCallum et al. 1999; Osborne & Costello 2004). A smaller sample size or respondent: item ratio may mean more unstable loadings and factors, random, non-replicable factors, and non-generalizable results (MacCallum et al. 1999; Osborne & Costello 2004). Sample size is, however, always constrained by resources available, and more often than not, scale development can be difficult to fund. (p. 8).

Making note of both the importance of number of items and the standard range of 200–300 for factor analysis, along with the proposed availability of cadets, we sought to collect data from at least 200 participants.

2.4 Procedure

Prior to data collection, an internal pilot study was conducted with our research team to determine if there is sufficient time to complete the tasks and to discover any problems with the procedure or the Qualtrics hosting platform.

Upon accessing the Qualtrics hosting platform's site, each participant read a description of the study objectives and then read and signed the consent form if they agreed to participate. They were then asked to complete the demographics page, gaming inventory, immersive tendencies questionnaire, and personality inventory. After they completed the pre-task questionnaires, they began the study. Participants were given a set of instructions to acquaint themselves with the task. No other training was required. During the study participants read through short, written vignettes that asked participants to imagine they were part of a HAT that was instructed to work together. They then viewed the short video clips (approximately 3 min or less) of the same nature. These video clips featured high or low cohesion teams consisting of human and robot team members performing a collaborative task that was specific to each of the subfactors of cohesion described previously. After watching each video, they were asked to rate their perceived level of the team's cohesion using the relevant items from our scale. This consisted of evaluating the team via the items in our item pool. Additionally, after each video clip participants were also asked to fill out the Group Cohesion Questionnaire. The study took no longer than 1.5 h. Upon completion of the study, participants were thanked for their participation and given course credit. Additionally, the PI's contact information was made available should they have any further questions or follow-up concerns.

2.5 Experimental Design

This study used a within-subjects design, where each participant was given the same vignettes and video clips to evaluate. Specific variables to be manipulated included vignettes of high and low cohesive teams to ensure that the items from our scale were indeed measuring cohesion. As such, rated items for the high and low cohesive teams should differ.

We also measured several covariates such as personality, existing attitudes towards robots, system trustworthiness, immersive tendencies, mood, and an existing cohesion scale to determine if any of these variables relate to the participants' ratings of team cohesion.

2.6 Data Analysis

We used confirmatory factor analysis (CFA) to examine whether the items appeared to be caused by the latent factors hypothesized during item development (e.g., subfactors of exclusivity, morale, complementarity). The primary metrics on which each subscale was evaluated were (1) *model fit*, (2) *internal consistency*, (3) *invariance*, (4) *sensitivity*, and (5) *relationship with existing criterion measures*.

Model fit represents the difference between model-implied response patterns and observed response patterns and is tested with a chi-square distribution comparing model-implied and observed variance-covariance matrices. When the test of model fit is significant, it indicates that an observed association is not modeled appropriately in the analysis.

We assessed *internal consistency*, which is one measure of scale reliability, using McDonald's ω. Although Cronbach's α is perhaps the most recognizable measure of internal consistency, it relies on a strict assumption that all items on a scale are equally relevant to the construct (i.e., all items have equal loadings), which is unrealistic in most cases (Dunn et al. 2014). In contrast, McDonald's ω allows each item to have a unique loading, which accurately represents how we modeled these data. Invariance represents how well the same model can represent the same scale across multiple occasions (e.g., across experimental conditions or over time).

We tested for configural, metric, and scalar *invariance*. Configural invariance means that the same factor structure is appropriate for all measurement occasions; metric invariance means the same factor structure and item loadings are appropriate for all occasions, and scalar invariance means that the same factor structure, item loadings, and item intercepts are appropriate for all occasions. Critically, scalar invariance is necessary for drawing unambiguous conclusions regarding mean differences on the latent factor—in other words, to know whether people perceived greater social cohesion when viewing "high" versus "low" cohesion scenarios, the scale must show scalar invariance between those two conditions, enabling us to test each scale's *sensitivity*. If any scale items fail to meet these levels of invariance, as determined by a significant change in a χ^2 test of model fit, this indicates that participants did not interpret the item meaning in the same way across conditions and could thus warrant item removal. Finally, we determined the scale's *relationship with existing criterion measures*. After fitting the best model we could for each subdimension, we used exploratory factor analysis (EFA) to test whether the items from these subdimensions were better modeled under one of the GEQ-10 factors or whether they formed distinct factors and, further, whether any distinct factors were correlated with the GEQ-10 factors.

Before examining the psychometric properties of each experimental subdimension, we focused on identifying a well-fitting model for the GEQ, as this was the only measure participants responded to in every scenario. We planned to begin by testing a three-factor congeneric model, where each item loaded solely on one of three factors: task cohesion, social cohesion, or attraction to the group. If this model did not fit the data well, then we removed items with the lowest factor loadings. If this congeneric model still did not fit the data well, then we used exploratory factor analysis to re-examine the model structure and construct a model that fit the data well for every measurement occasion, to the best of our ability.

To downselect items from the larger pool, we used a multi-stage process. First, we tested each of the theorized subdimensions as single-factor models. In some cases, factors were only informed by two indicators, which are under-identified for factor analysis, so we tested these factors in conjunction with another factor that was assessed at the same measurement occasion. At this stage, if the single-factor model did not fit the data well, we identified items with the smallest factor loadings to be removed, from lowest to highest. In cases where the factor was only informed by three indicators, which produces a saturated model that cannot be tested for model fit, we targeted items for removal with standardized loadings below .5. Our goal at this stage was to find a model that had adequate fit in both high and low cohesion scenarios, using the exact same items in both scenarios, until there were no fewer than three items per factor.

Second, we tested these models for configural, metric, and scalar invariance. To test for configural invariance, we correlated the single-factor models for high and low cohesion—if the configural invariant model did not fit the data, then we determined the scale was not invariant and that the items making up the scale were suspect; on the other hand, if the model did fit the data well, we tested for metric invariance. To test for metric invariance, we constrained item loadings in the high cohesion scenario to be identical to item loadings in the low cohesion scenario. At this stage, if the full scale was not metric invariant, then we tested for partial invariance by constraining each item individually—any items that were not metric invariant were marked for possible removal, whereas any items that were metric invariant were tested for scalar invariance. To test for scalar invariance, we constrained all metric invariant items to have identical item intercepts to each other. At this stage, we also had to freely estimate one of the latent factor means, otherwise the scalar invariant model would always fit much worse than the metric invariant model—each time, we freely estimated the latent mean for responses to the low cohesion scenario. Here, any items that demonstrated scalar invariance were marked for inclusion, whereas all others were marked for possible removal. Third, and finally, we conducted exploratory factor analysis using the GEQ-10 along with

each subdimension measured along with that instance of the GEQ. This process will tell us whether the new subdimensions' items are better represented as part of one of the GEQ-10 subfactors or whether it is better modeled as a distinct factor—any items from the experimental subdimensions that cluster with either of the GEQ-10 subfactors would be marked for removal.

In addition to downselecting items, we also had the goal of having at least enough items to test each subdimension with the GEQ-10 in the third stage, meaning that if any subdimensions were not invariant, we still conducted exploratory factor analysis of the best-fitting model for the subdimension with our best-fitting model of the GEQ-10.

All factor analyses for our cohesion measures were conducted in R v. 4.0.2 (R Core Team 2020). Confirmatory factor analyses were conducted using the "lavaan" package (Rosseel 2012), whereas internal consistency measures and exploratory analyses were conducted using the "psych" package (Revelle 2022).

3. Results

Participants' scores on the Video Game Experience (VGE) scale, Immersive Tendencies Questionnaires (ITQ), and Negative Attitude Towards Robots Scales (NARS) were calculated; these scores were then correlated with an aggregation of participants' ratings of the team's cohesion across all vignettes (i.e., total aggregate cohesion), across high cohesion teams' vignettes (i.e., aggregate "positive" cohesion), and across low cohesion teams' vignettes (i.e., aggregate "low" cohesion). The results of these scales did not have significant relationships to subjective ratings of cohesion, nor did they affect scale validation efforts. The specific analyses and details regarding said analyses are available: VGE information is available in Appendix B; ITQ information is available in Appendix C; NARS information is available in Appendix D.

The same analyses were done with the "Familiarity with the Clone Wars television show," question, PANAS, and Mini-IPIP scales. However, these analyses were relevant to the subjective ratings of cohesion and/or the scale validation effort. These analyses are detailed in the following sections.

3.1 Demographics, Individual Differences, and Subjective Cohesion Ratings

The first set of analyses focused on analyzing demographics and pre-task data to understand the current sample population as well as any relationships between the subjective data and reported cohesion ratings for both high and low cohesion

scenarios. The following section will present and outline the demographic and individual difference data, followed by the CFA results.

3.1.1 Familiarity with *Star Wars Clone Wars* Series

Prior to the experiment, participants were asked to report how familiar, on a scale of 1 (not at all) to 5 (very familiar) they were with the *Star Wars Clone Wars* television series. It was not expected that participants would be overly familiar with the television series, as that may impact their perceived relationship to the teams depicted in the vignettes. Indeed, results indicated that mean familiarity was generally moderate to low as indicated by the mean response rate (M = 2.88, SD = 1.40) for our sample. Table 1 reports the overall familiarity scores for each rating.

Table 1 Familiarity with *Star Wars Clone Wars* series. Note: N= 282 as 13 participants chose not to respond to this question.

Self-reported score	Familiarity
1	55 (19.50%)
2	76 (26.95%)
3	53 (18.79%)
4	44 (15.60%)
5	54 (19.15%)

3.1.2 Positive and Negative Attitudes Scale (PANAS-X) Overall Statistics and Subdimensions

The PANAS-X measures both Positive Affect (PA) and Negative Affect (NA), see Fig. 3, as well as 11 primary affects labeled: Fear, Sadness, Guilt, Hostility, Shyness, Fatigue, Surprise, Joviality, Self-Assurance, Attentiveness, and Serenity (Fig. 4). Results indicated that PA was self-reported as being higher prior to the start of the experiment, compared to NA.

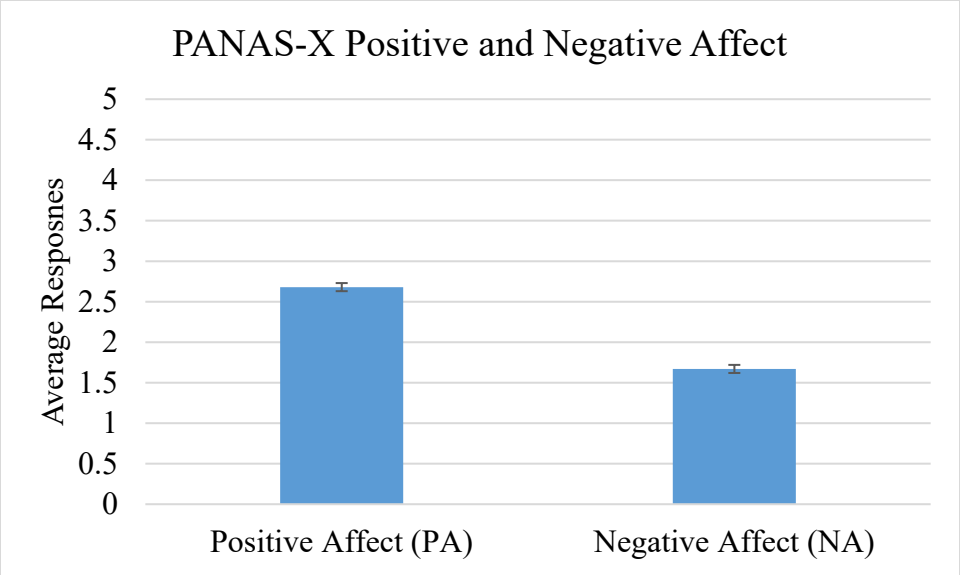

Fig. 3 Averaged responses for the PANAS-X's positive and negative affect subscales. The minimum and maximum scores are 1 and 5, respectively.

Additionally, it appears that the subdimensions of PA (i.e., Joviality, Self-Assurance, Attentiveness) were also slightly higher to begin with compared to those emotions relating to NA (i.e., Fear, Hostility, Guilt, Sadness).

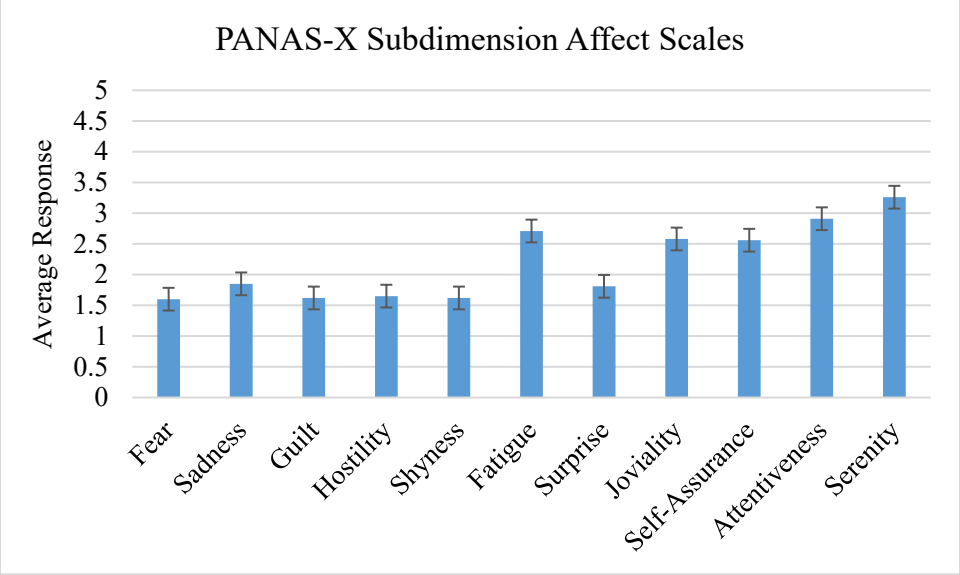

Fig. 4 Averaged responses for the PANAS-X's subscales. The minimum and maximum scores are 1 and 5.

Additional analyses were performed between the PANAS-X data and the aggregated cohesion score for the GEQ-10 responses. Two aggregate cohesion scores were calculated as the average responses from the GEQ-10 following the presentation of both the high (i.e., aggregate "positive" cohesion) and low cohesion

video clips (i.e., aggregate "negative" cohesion). A third aggregate cohesion score, Total Aggregate cohesion—the average of GEQ scores following presentation of all cohesion video clips—was calculated. We performed Kolmogorov-Smirnov tests on all three aggregate cohesion scores to ascertain normality and found that total aggregate cohesion ($D(293) = 0.173$, $p<0.001$), aggregate positive cohesion ($D(293) = 0.159$, $p<0.001$), and aggregate negative cohesion ($D(293) = 0.119$, $p<0.001$) were all significant, indicating the data distribution was non-normal.

Consequently, Spearman's Rho was used to determine if any significant relationships existed between the aggregate cohesion scores and the subjective response and individual difference information. Bivariate correlations are presented in Tables 2 and 3 and revealed a positive correlation between responses on the PANAS-X for general PA score and the aggregate "positive" cohesion score—the average of GEQ-10 scores following presentation of the "positive" or high cohesion video clips. In other words, those who reported higher PA on the PANAS-X, prior to the start of the experiment, also rated cohesion higher following presentation of the high cohesive video clips.

Table 2 Correlations (Spearman's Rho) between self-reported PA and aggregate cohesion

	Total aggregate cohesion	Aggregate "positive" cohesion	Aggregate "negative" cohesion
PANAS-X (positive affect)	**0.163**[a]	**0.130**[b]	0.066

[a] Correlation is significant at the 0.01 level (2-tailed).

[b] Correlation is significant at the 0.05 level.

A similar relationship was found for the responses relating to NA on the PANAS-X and the aggregate "negative" cohesion score—the average of GEQ-10 scores following presentation of low or "negative" cohesion video clips . Here a significant relationship was found and indicates that those individuals who self-reported high NA prior to the start of the experiment also rated cohesion higher following presentation of the low cohesion video clips. However, those who self-reported high NA prior to the start of the experiment also reported higher cohesion following "positive" cohesion video clips.

Table 3 Correlations (Spearman's Rho) between self-reported NA and aggregate cohesion

	Total aggregate cohesion	Aggregate "positive" cohesion	Aggregate "negative" cohesion
PANAS-X (negative affect)	**0.299**[a]	**0.357** [a]	**0.128**[b]

[a] Correlation is significant at the 0.01 level (2-tailed).

[b] Correlation is significant at the 0.05 level.

3.1.3 Mini-IPIP Personality Inventory Overall Statistics and Subdimensions

The Mini-IPIP personality assessment was also used to measure the Big Five personality traits: Agreeableness, Extraversion, Intellect, Conscientiousness, and Neuroticism. Figure 5 illustrates the personality ratings for each of these five subdimensions.

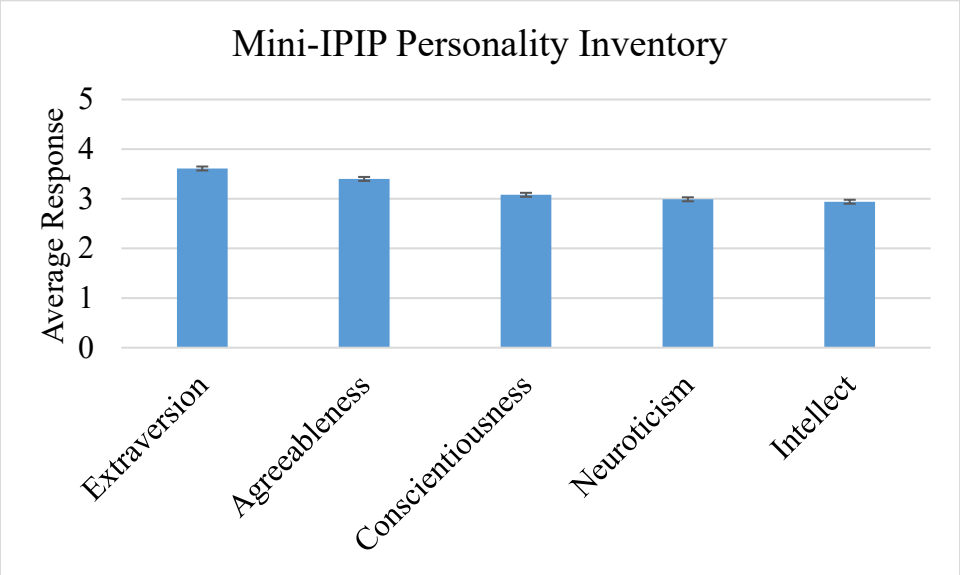

Fig. 5 Participants' averaged scores for the "Big Five" personality factors. The minimum and maximum scores are 1 and 5.

Like the previous analyses, correlations were also performed between the aggregate cohesion scores and the Big Five personality dimensions. As shown in Table 4, the only significant relationships found were a negative relationship between *extraversion* and aggregate "positive" cohesion, a negative relationship between agreeableness and aggregate "positive" cohesion, and a negative relationship between agreeableness and total aggregate cohesion.

Table 4 Summary of the correlation between total, "positive," and "negative" cohesion and Big Five personality factors. All values reported are Spearman's Rho.

Personality Factor	Total aggregate cohesion	Aggregate "positive" cohesion	Aggregate "negative" cohesion
Mini-IPIP (extraversion)	–.0103	**–0.172**[a]	0.017
Mini-IPIP (agreeableness)	**–0.123**[b]	**–0.145**[b]	–0.009
Mini-IPIP (conscientiousness)	0.039	0.003	0.075
Mini-IPIP (neuroticism)	0.034	0.056	0.018
Mini-IPIP (intellect)	0.005	0.067	–100.071

[a] Correlation is significant at the .01 level (2-tailed).

[b] Correlation is significant at the .05 level.

3.2 Cohesion Scale Evaluation

We tested model goodness of fit on several CFA models, twice for each distinct subscale (once for responses to depictions of high cohesion and once for low cohesion)—except for the GEQ-10, which was administered on 18 occasions. Because the GEQ-10 represented our criterion measure of team cohesion, we prioritized testing whether the factor structure posited by Carless and De Paola (2000) was able to adequately fit the data we obtained here. If the GEQ model did not fit well, then we would have to reassess to develop a well-fitting GEQ model so we could fully evaluate our other items. Items from the GEQ-10 are presented in Appendix E.

A high-level summary of the factor analyses is presented in Table 5. The table conveys which items were retained for each subdimension, the internal consistency of each subdimension, whether the items were invariant, and the effect size of evaluating high versus low cohesion scenarios (only for invariant scales). More specifically, the "items included" column represents those items with the highest factor loadings that also produced a well-fitting model—in most cases, including more than three items resulted in poor fit, whereas including only three items produced models with estimable parameters but untestable model fit; for one construct (team flexibility; items 71 and 72), we had just two items, meaning we could not model the construct by itself, let alone test it for goodness of fit. The final "items retained" column includes only those nine items that demonstrated excellent psychometric properties across all tests.

Table 5 Summary of item reduction results

Factor	Items included	Model fit	Int. consist. (high;low)	Invariance			ES	Items retained
				C	M	S		
GEQ – Task	2,3,4	NT	. . .	. . .	. . .	. . .	. . .	. . .
GEQ – Social	6,7,8	NT	. . .	. . .	. . .	. . .	. . .	. . .
GEQ – Attraction	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Function	1,3,4	NT	0.93; 0.90	✓	X	X	. . .	. . .
Exclusivity	15,16,17	NT	0.88; 0.95	✓	X	X	. . .	. . .
Complementarity	55,57,61,62	✓	0.90; 0.85	✓	✓	✓	1.353	55,57,61,62
Pride	30,31,32	NT	0.90; 0.82	X	X	X	. . .	. . .
Morale	52,53,54	NT	0.74; 0.80	✓	✓	✓	1.607	52,53,54
Belongingness	48,49,50	NT	0.79; 0.83	X	X	X	. . .	. . .
Attraction to the Group	20,21,22	NT	0.85; 0.83	X	X	X	. . .	. . .
Social	38,39,46	NT	0.95; 0.95	✓	X	X	. . .	. . .
Leadership Direction	26,28,29	NT	0.91; 0.95	✓	✓[a]	✓[a]	2.112	26
Team Learning Orientation	66,67,70	NT	0.84; 0.93	X	X	X	. . .	. . .
Shared Language	73,74,75	NT	0.85; 0.85	✓	✓[a]	X	. . .	. . .
Team Flexibility	71,72	NT	. . .	X	X	X	. . .	. . .
Perceived Efficacy	78,79,81	NT	0.90; 0.92	✓	✓[a]	✓[a]	1.509	81

[a] Partial invariance.

Note: NT = not testable. Internal consistency measures were computed using McDonald's ω for high and low cohesion scenarios, respectively. Invariance columns indicate whether the model met configural (C), Metric (M), and Scalar (S) invariance. Effect sizes (ESs) are differences between latent means for high and low cohesion scenarios in terms of standard deviations.

Additionally, to better understand the process taken for testing invariance between item responses to the low and high cohesion scenarios, a path diagram of our modeling approach is depicted in Fig. 6. Here, the large circles represent each latent construct we are measuring (morale, belongingness, etc.), which are informed by the common variance shared among item responses. The curved line connecting the construct across scenarios represents the covariance (φ) between measurements. In our specification, we scaled the latent variables using a mean (triangles) of 0 and variance (self-directed curved line) of 1—if we met scalar invariance, we freely estimated the mean of the construct for high cohesion scenarios (hence, 0*). The relationship between the construct and each item response is represented by λ—when testing metric invariance, λs for corresponding items across scenarios are constrained equal. The intercepts (triangles) for each item response are depicted with βs—when testing scalar invariance, βs for corresponding items are constrained equal. Error terms are represented with εs and their variances with θs. We estimated error variances but did not test for their invariance across scenarios.

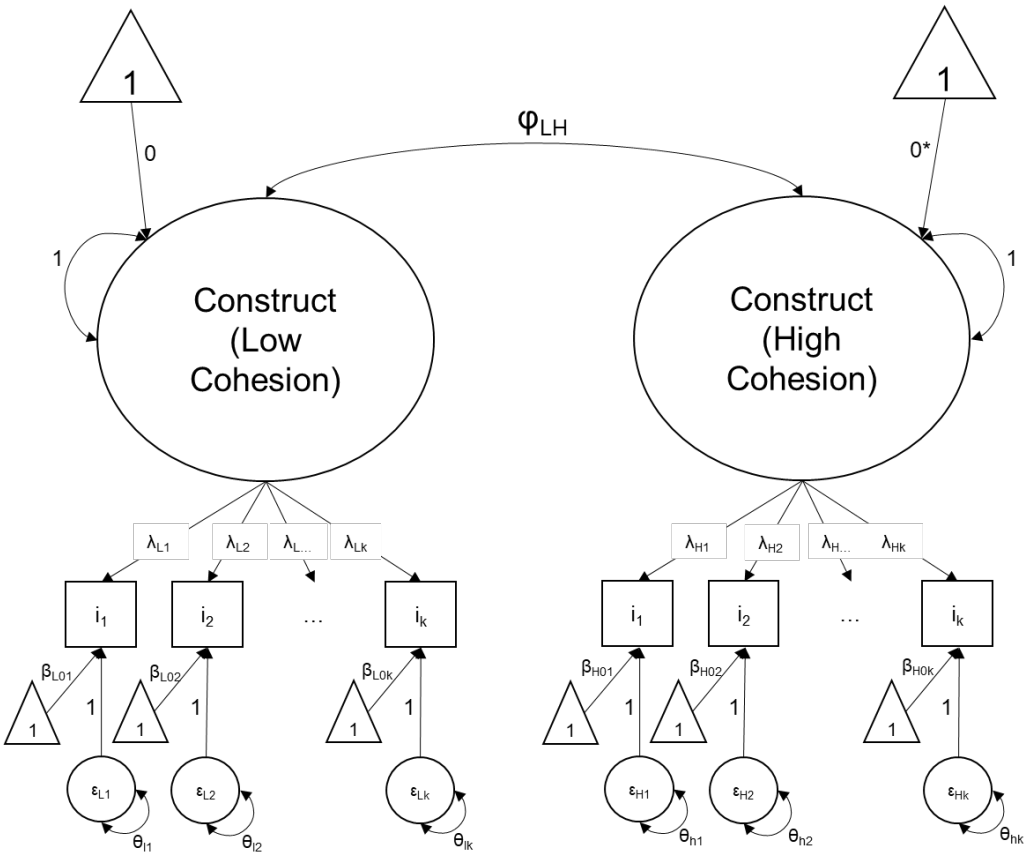

Fig. 6 Structural equation model representing our approach to testing invariance between item responses to the low and high cohesion scenarios

3.2.1 GEQ

First, we tested a model wherein items on the GEQ were caused by three factors: task cohesion (items 1–4), social cohesion (items 5–8), and attraction to the group (items 9 and 10). Ideally, each of these subfactors would be tested individually for local fit before testing in the larger model context—however, the attraction to the group factor is only informed by two items, which does not provide enough information to fit a model without also including other factors, so we modeled all three factors concurrently. In this initial model, each item only loaded onto one subfactor, and we tested this factor structure on each occasion. Our results indicated that the published three-factor model of the GEQ did not fit the data well on any of our measurement occasions, thus indicating some major problems with the factor structure, which has also been found in previous work (Carron and Brawley 2000).

At this stage, we decided to conduct exploratory factor analysis, extracting three factors using geomin rotation, which allows factors to correlate. Here, only one measurement occasion exhibited good fit, though none of these exploratory models clustered items according to Carless and De Paola's (2000) factor structure. With few exceptions, one factor was made up of one social cohesion item, one task cohesion item, and both attraction to the group items, while the other items designed to measure task and social cohesion generally clustered with their respective factors. Unfortunately, there was no obvious connection between the content of the attraction to the group items and the other items it clustered with; therefore, we decided to remove the attraction to the group subfactor and test the remaining subfactors in a two-factor exploratory analysis.

The two-factor exploratory model revealed many inconsistencies with the same items that typically loaded with the attraction to the group items—they often had very high cross-loadings between the two factors. Additionally, the two-factor exploratory model still had poor fit for all but one measurement occasion. Therefore, we opted to remove one item from the task cohesion factor, "Our team is united in trying to reach its goals for performance," and one item from the social cohesion subfactor, "Our team would like to spend time together outside of work hours," leaving us with three items to measure task cohesion and three items to measure social cohesion.

To be sure this two-factor structure worked well for all measurement occasions; we conducted another set of two-factor exploratory analyses and found that we had acceptable model fit for all but two measurement occasions, and that the task and social cohesion items loaded on their respective factors for all but one measurement occasion. In all subsequent analyses, we modeled the GEQ with two factors, one

representing task cohesion and one representing social cohesion, and allowed items to cross-load on both factors.

Ideally, at this stage we would test the model for invariance. However, because we measured the GEQ on 18 occasions, thorough invariance checking would be untenable given our relatively limited sample size. Furthermore, the purpose of this study was not to test the properties of the GEQ, but to use the GEQ to test certain properties of several experimental items that might be useful for measuring other dimensions beyond the GEQ. Specifically, we used the GEQ to examine whether these non-GEQ items are better modeled under the existing GEQ factors or better modeled as distinct factors and, if modeled as distinct factors, whether these factors correlated with the GEQ factors.

3.2.2 Function

For the function-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 45 and as many as 96 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our function-based cohesion items did not fit the data well for high, $\chi^2(35) = 367.497$, $p < 0.001$, or low cohesion scenarios, $\chi^2(35) = 241.895$, $p < 0.001$. No model fit the data well when there were more than three indicators and the three indicators with the highest factor loadings were "Our team was committed to working together" ($\lambda_{high} = 0.875$; $\lambda_{low} = 0.866$), "Our team accomplished assigned tasks to the best of their ability" ($\lambda_{high} = 0.912$; $\lambda_{low} = 0.839$), and "Our team was united in working toward task goals" ($\lambda_{high} = .916$; $\lambda_{low} = 0.897$). Internal consistency was excellent for high and low cohesion scenarios, McDonald's $\omega = 0.93$ and McDonald's $\omega = 0.90$, respectively, indicating that these three items reliably reflect the same construct. With only three indicators, our model of function-based cohesion was saturated, thereby making goodness of fit untestable.

In our tests of invariance, our configural-invariant model fit the data well, $\chi^2(8) = 12.627$, $p = 0.125$; however, our metric-invariant model did not, $\chi^2(11) = 35.077$, $p < 0.001$. Even after testing each item separately for metric invariance, we could not achieve good model fit. These results indicate that the items designed to measure function-based cohesion may be too problematic to include in future measures of cohesion.

Next, we examined how these three items clustered in an exploratory analysis with the six items from the GEQ. In a two-factor EFA, we observed very poor model fit for both high, $\chi^2(19) = 140.64$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 104.418$, $p < 0.001$. In a three-factor EFA, we still had poor fit for the high, $\chi^2(12) = 29.585$, $p = 0.003$, and low cohesion scenarios, $\chi^2(12) = 34.664$, $p = 0.001$, but the fit was significantly better for the three- versus two-factor models. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and function items into distinct factors. In the high cohesion scenario, we observed small-medium correlations between the function factor and GEQ task, $r = 0.41$, and GEQ social, $r = 0.34$; in the low cohesion scenario, these respective correlations were small, $r = .024$ and $r = 0.22$.

Overall, these results suggest that if the function items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the function-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that respondents may not understand these statements to mean the same thing in different scenarios.

3.2.3 Exclusivity

For the exclusivity-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 51 and as many as 113 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our exclusivity-based cohesion items did not fit the data well for high, $\chi^2(27) = 249.642$, $p < 0.001$, or low cohesion scenarios, $\chi^2(27) = 221.564$, $p < 0.001$. No model fit the data well when there were more than three indicators and the three indicators with the highest factor loadings were "This team has clearly established norms for working together" ($\lambda_{high} = 0.764$; $\lambda_{low} = 0.892$), "Most of the team members fit what I feel to be the idea of a good teammate" ($\lambda_{high} = 0.890$; $\lambda_{low} = 0.954$), and "Team members embody the ideal for a good team member" ($\lambda_{high} = 0.859$; $\lambda_{low} = 0.941$). Internal consistency was excellent for high and low cohesion scenarios, McDonald's $\omega = 0.88$ and McDonald's $\omega = 0.95$, respectively, indicating that these three items reliably reflect the same construct. With only three indicators, our model of exclusivity-based cohesion was saturated, thus making goodness of fit untestable.

Our configural-invariant model fit the data well, $\chi^2(8) = 14.217$, $p = 0.076$; however, our metric-invariant model did not, $\chi^2(11) = 66.286$, $p < 0.001$. After testing each item separately, we found none of them to be metric invariant, indicating that these items do not appear to have desirable enough psychometric properties to be included in future measures.

In the two-factor EFA, we observed very poor model fit for both high, $\chi^2(19) = 75.441$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 108.923$, $p < 0.001$. The three-factor exploratory model did fit well for both the high, $\chi^2(12) = 16.763$, $p = 0.159$, and low cohesion scenarios, $\chi^2(12) = 15.814$, $p = 0.200$. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and exclusivity items into distinct factors. In the high cohesion scenario, we observed small-medium correlations between the exclusivity factor and GEQ task, $r = 0.40$, and GEQ social, $r = 0.32$, as well as in the low cohesion scenario, $r = 0.32$ and $r = 0.33$, respectively.

Overall, these results suggest that if the exclusivity items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the exclusivity-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that respondents may not understand these statements to mean the same thing in different scenarios.

3.2.4 Complementarity

For the complementarity-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 44 and as many as 107 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our team complementarity items did not fit the data well for high, $\chi^2(44) = 166.118$, $p < 0.001$, or low cohesion scenarios, $\chi^2(44) = 201.301$, $p < 0.001$. After removing all but four items, our single-factor model fit the data for both high, $\chi^2(2) = 1.394$, $p = 0.498$, and low cohesion scenarios, $\chi^2(2) = 5.541$, $p = 0.063$. The four indicators in this model were "Individual members of the team are important because they offer skills and abilities that work well together" ($\lambda_{high} = 0.833$; $\lambda_{low} = 0.741$), "My teammates rely on me because I have skills that they do not have" ($\lambda_{high} = 0.760$; $\lambda_{low} = 0.653$), "The skills of the autonomous teammate(s) complement me in things I am not good at" ($\lambda_{high} = 0.847$; $\lambda_{low} = 0.799$), and "The other members and I compensate for each other's weaknesses" ($\lambda_{high} = 0.897$; λ_{low}

= 0.845). Internal consistency was excellent for high and low cohesion scenarios, McDonald's ω = 0.90 and McDonald's ω = 0.85, respectively, indicating that these four items reliably reflect the same construct.

Our configural-invariant model fit the data well, $\chi^2(19) = 14.580$, $p = 0.749$, as did our metric-invariant model, $\chi^2(23) = 24.987$, $p = 0.351$, and our scalar invariant model, $\chi^2(29) = 37.864$, $p = 0.125$. These results indicate that our measure of complementarity has desirable psychometric properties and that these four items would be worth including in future measures.

Because our measure of complementarity was scalar invariant, we could perform a meaningful test of the mean differences between high and low cohesion scenarios, which indicated that participants scored 1.353 standard deviations lower on complementarity when viewing low cohesion scenarios compared to high cohesion scenarios, $z = -12.345$, $p < 0.001$.

A two-factor exploratory analysis produced very poor model fit for both high, $\chi^2(26) = 216.567$, $p < 0.001$, and low cohesion scenarios, $\chi^2(26) = 119.314$, $p < 0.001$. Although the three-factor exploratory model did not fit well for the high cohesion scenario, $\chi^2(18) = 41.577$, $p = 0.001$, it did fit well for the low cohesion scenarios, $\chi^2(18) = 20.120$, $p = 0.326$. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and complementarity items into distinct factors. In the high cohesion scenario, we observed small correlations between the complementarity factor and GEQ task, $r = 0.32$, and GEQ social, $r = 0.23$, as well as in the low cohesion scenario, $r = 0.31$ and $r = 0.22$, respectively.

Overall, these results suggest that the four-item measure of complementarity has good psychometric properties of strong internal consistency, scalar invariance, sensitivity to depictions of high and low cohesion, and is both distinct from, and correlated with, task and social cohesion.

3.2.5 Pride

For the pride-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 61 and as many as 73 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

Our measure of pride contained only three items, which is enough to fit a single-factor model but not enough to test the model's goodness of fit. Internal consistency was very good for high and low cohesion scenarios, McDonald's ω = 0.90 and

McDonald's ω = 0.82, respectively, indicating that these three items reliably reflect the same construct.

Our configural-invariant model did not fit the data well, $\chi^2(8) = 21.786$, $p = 0.005$, so we could not test for metric or scalar invariance, indicating that these items do not appear to have desirable enough psychometric properties to be included in future measures.

In a two-factor exploratory analysis, we observed very poor model fit for both high, $\chi^2(19) = 141.003$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 67.283$, $p < 0.001$. However, the three-factor exploratory model fit well for both the high, $\chi^2(12) = 9.97$, $p = 0.619$, and low cohesion scenarios, $\chi^2(12) = 17.202$, $p = 0.142$. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and pride items into distinct factors. In the high cohesion scenario, we observed small-medium correlations between the pride factor and GEQ task, $r = 0.39$, and GEQ social, $r = 0.35$, and small correlations in the low cohesion scenario, $r = 0.19$ and $r = 0.27$, respectively.

Overall, these results suggest that if the pride items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the pride-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that respondents may not understand these statements about team pride to mean the same thing in different contexts.

3.2.6 Morale

For the morale-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 33 and as many as 78 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our morale-based cohesion items fit the data well for high, $\chi^2(2) = 1.576$, $p = 0.455$, and low cohesion scenarios, $\chi^2(2) = 4.478$, $p = 0.107$. Internal consistency was adequate for high and low cohesion scenarios, McDonald's ω = 0.74 and McDonald's ω = 0.80, respectively, indicating that these four items reliably reflect the same construct.

Our configural-invariant model did not fit the data well, $\chi^2(19) = 36.007$, $p = 0.011$. We could still afford to remove one item at this stage, so we removed the item with the lowest factor loading in both scenarios, which was "Working with my team

makes me feel good" ($\lambda_{high} = 0.531$; $\lambda_{low} = 0.517$). After removing this item, internal consistency remained virtually identical for both high and low cohesion scenarios, McDonald's $\omega = 0.73$ and McDonald's $\omega = 0.81$, respectively. After retesting for invariance, we found the configural invariant model fit well, $\chi^2(8) = 9.940$, $p = 0.269$, as did the metric invariant model, $\chi^2(11) = 12.952$, $p = 0.296$, and the scalar invariant model, $\chi^2(13) = 13.338$, $p = 0.422$. These results indicate that our measure of team morale has desirable psychometric properties and that these three items would be worth including in future measures.

Because our measure of team morale was scalar invariant, we could perform a meaningful test of the mean differences between high and low cohesion scenarios, which indicated that participants scored 1.607 standard deviations lower on morale when viewing low cohesion scenarios compared to high cohesion scenarios, $z = -11.928$, $p < 0.001$.

We then conducted a two-factor exploratory analysis, which produced poor model fit for both high, $\chi^2(19) = 34.971$, $p = 0.014$, and low cohesion scenarios, $\chi^2(19) = 66.879$, $p < 0.001$. A three-factor exploratory model fit well for the high cohesion scenario, $\chi^2(12) = 10.385$, $p = 0.582$, but did not fit well for the low cohesion scenario, $\chi^2(12) = 21.855$, $p = 0.039$. In both scenarios, the three-factor model cleanly separated the GEQ task and GEQ social items; however, in the high scenario, some morale items had roughly equal cross-loadings, a morale subfactor and the GEQ task subfactor, whereas in the low cohesion scenario, morale items clearly loaded on a distinct factor. This may indicate that the morale items are distinct from the GEQ task and social subscales while also having substantial overlap with GEQ task cohesion. In the high cohesion scenario, we observed medium correlations between the morale factor and GEQ task, $r = 0.49$, and GEQ social, $r = .42$, and small-medium correlations in the low cohesion scenario, $r = 0.48$ and $r = 0.36$, respectively.

Overall, these results suggest that the three-item measure of team morale has desirable psychometric properties of good internal consistency, scalar invariance, sensitivity to depictions of high and low cohesion, and is both meaningfully distinct from, and correlated with, task and social cohesion.

3.2.7 Belongingness

For the belongingness-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 46 and as many as 88 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were

identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

Our measure of belongingness contained only three items, which is enough to fit a single-factor model but not enough to test the model's goodness of fit. Internal consistency was good for high and low cohesion scenarios, McDonald's $\omega = 0.79$ and McDonald's $\omega = 0.83$, respectively, indicating that these three items reliably reflect the same construct.

Next, we tested our belongingness items for invariance between the high and low cohesion scenarios. Our configural-invariant model did not fit the data well, $\chi^2(8) = 17.536$, $p = 0.025$, so we could not test for metric or scalar invariance, indicating that these items do not appear to have desirable enough psychometric properties to be included in future measures.

In a two-factor exploratory analysis, we observed very poor model fit for the high cohesion scenario, $\chi^2(19) = 123.249$, $p < 0.001$, but good model fit for the low cohesion scenario, $\chi^2(19) = 25.063$, $p = 0.158$. The three-factor exploratory model fit well for both the high, $\chi^2(12) = 14.045$, $p = 0.298$, and low cohesion scenarios, $\chi^2(12) = 8.271$, $p = 0.764$. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and belongingness items into distinct factors. In the high cohesion scenario, we observed small-medium correlations between the belongingness factor and GEQ task, $r = 0.42$, and GEQ social, $r = 0.39$, as well as in the low cohesion scenario, $r = 0.53$ and $r = 0.34$, respectively.

Overall, these results suggest that if the belongingness items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the belongingness-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that respondents may not understand these statements about team belongingness to retain the same meaning in different contexts.

3.2.8 Attraction to the Group

For the attraction to the group items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 33 and as many as 76 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our attraction to the group cohesion items did not fit the data well for high, $\chi^2(2) = 9.246$, $p = 0.01$, or low cohesion scenarios, $\chi^2(2) =$

10.903, $p = 0.004$. We then removed the item with the lowest factor loading, which was, "If given the chance, my teammates would have me leave the team and join another" ($\lambda_{high} = 0.758$; $\lambda_{low} = 0.427$). Internal consistency was good for high and low cohesion scenarios, McDonald's $\omega = 0.85$ and McDonald's $\omega = 0.83$, respectively, indicating that these four items reliably reflect the same construct. With only three items, the model was saturated, and we could not meaningfully test goodness of fit.

Our configural-invariant model did not fit the data well, $\chi^2(8) = 40.283$, $p < 0.001$. These results indicate that our measure of team attraction to the group may be too problematic to include in future measures.

We then conducted a two-factor exploratory analysis, which produced poor model fit for both high, $\chi^2(19) = 69.345$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 97.207$, $p < 0.001$. The three-factor exploratory model fit well for the high cohesion scenario, $\chi^2(12) = 17.107$, $p = 0.146$, but did not fit well for the low cohesion scenario, $\chi^2(12) = 12.757$, $p = 0.387$. In both scenarios, the three-factor model separated GEQ task, GEQ social, and attraction to the group items into distinct factors, although in the high cohesion scenario, one task cohesion item had strong cross-loadings with social cohesion, loading higher onto the social cohesion factor. In the high cohesion scenario, we observed small-medium correlations between the morale factor and GEQ task, $r = 0.35$, and GEQ social, $r = 0.40$, and small-medium correlations in the low cohesion scenario, $r = 0.41$ and $r = 0.39$, respectively.

Overall, these results suggest that if the attraction to the group items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the attraction-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that respondents may not interpret these statements about individual attraction to the group in the same way in different contexts.

3.2.9 Social

For the social cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 44 and as many as 62 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our social cohesion items did not fit the data well for high, $\chi^2(90) = 423.108$, $p < 0.001$, or low cohesion scenarios, $\chi^2(90) = 474.797$, $p < 0.001$. No model fit the data well when there were more than three indicators and

the three indicators with the highest factor loadings were "I liked the team I was in" ($\lambda_{high} = 0.955$; $\lambda_{low} = 0.952$), "I enjoyed interacting with this team" ($\lambda_{high} = 0.932$; $\lambda_{low} = 0.951$), and "Members in this team respect one another" ($\lambda_{high} = 0.913$; $\lambda_{low} = 0.898$). Internal consistency was excellent for high and low cohesion scenarios, McDonald's $\omega = 0.95$ and McDonald's $\omega = 0.95$, respectively, indicating that these three items reliably reflect the same construct. With only three indicators, our model of function-based cohesion was saturated, thereby making goodness of fit untestable.

Our configural-invariant model fit the data well, $\chi^2(8) = 6.618$, $p = 0.578$; however, our metric-invariant model did not, $\chi^2(11) = 44.875$, $p < 0.001$. Even after testing each item separately for metric invariance, we could not achieve good model fit. These results indicate that the items designed to measure social cohesion may be too problematic to include in future measures of cohesion.

A two-factor exploratory factor analysis to assess whether the function items clustered under the same factor as either task or social cohesion. We observed very poor model fit for both high, $\chi^2(19) = 141.003$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 67.283$, $p < 0.001$. In a three-factor exploratory analysis, the model exhibited good fit for both high, $\chi^2(12) = 14.866$, $p = 0.249$, and low cohesion scenarios, $\chi^2(12) = 12.028$, $p = 0.443$. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and non-GEQ social items into distinct factors. In the high cohesion scenario, we observed small-medium correlations between the function factor and GEQ task, $r = 0.40$, and GEQ social, $r = 0.35$, as well as in the low cohesion scenario, $r = 0.37$ and $r = 0.32$, respectively.

These results raise major concerns because the social cohesion items from the non-GEQ scale did not cluster with the social cohesion items from the GEQ, nor did the non-GEQ social cohesion scale correlate more strongly with the GEQ social cohesion factor than with the GEQ task cohesion factor. We may be tempted to attribute these problems to the issues we found with fitting the GEQ (see Section 3.3.1), which may suggest that the GEQ was a poor choice as our criterion measure of team cohesion. Still, the non-GEQ measure of social cohesion was not invariant, thus calling into question whether either measure of social cohesion is adequate.

3.2.10 Leadership Direction

For the leadership direction-based cohesion items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 55 and as many as 82 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations

were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our leadership direction items did not fit the data well for high, $\chi^2(9) = 55.081$, $p < 0.001$, or low cohesion scenarios, $\chi^2(9) = 33.738$, $p < 0.001$. After removing the two items with the lowest factor loadings, we had a four-item measure that fit the data well in both high, $\chi^2(2) = 1.165$, $p = 0.559$, and low cohesion scenarios, $\chi^2(2) = 0.089$, $p = 0.957$. The four items used for this model were "When an individual in this team needs help, their leaders acknowledge those needs and their importance" ($\lambda_{high} = 0.890$; $\lambda_{low} = 0.862$), "Leaders understand the capabilities of this team" ($\lambda_{high} = 0.858$; $\lambda_{low} = 0.678$), "The leaders support team members during task performance" ($\lambda_{high} = 0.913$; $\lambda_{low} = 0.653$), and "The leaders work along with their team members" ($\lambda_{high} = 0.842$; $\lambda_{low} = 0.823$). Internal consistency was excellent for both high and low cohesion scenarios, McDonald's $\omega = 0.93$ and McDonald's $\omega = 0.89$, respectively.

Our configural-invariant model did not fit the data well, $\chi^2(19) = 44.311$, $p = 0.001$. We were able to achieve good fit for the configural invariant model by removing the item, "When an individual in this team needs help, their leaders acknowledge those needs and their importance." After doing so, the configural invariant model fit well, $\chi^2(8) = 14.000$, $p = 0.082$. The resulting model also had good internal consistency for both high and low cohesion scenarios, McDonald's $\omega = 0.91$ and McDonald's $\omega = 0.85$. We then tested the model for metric invariance, which did not have good fit, $\chi^2(11) = 46.316$, $p < 0.001$. After testing items individually, we found that the item, "Leaders understand the capabilities of this team," met criteria for both metric, $\chi^2(9) = 14.221$, $p = 0.115$, and scalar invariance, $\chi^2(9) = 14.221$, $p = 0.115$—these two models had identical fit because once the item's intercepts were constrained, we could freely estimate the latent mean of one factor, thus resulting in identical degrees of freedom and model fit. Indeed, freeing the latent mean for the low cohesion scenario revealed that participants rated leadership direction 2.112 standard deviations lower in the low versus high cohesion scenario, $z = -12.384$, $p < 0.001$.

A two-factor exploratory analysis produced very poor model fit for both high, $\chi^2(19) = 112.672$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 127.181$, $p < 0.001$. The three-factor exploratory model fit well for the high, $\chi^2(12) = 15.620$, $p = 0.209$, and low cohesion scenarios, $\chi^2(18) = 20.120$, $p = 0.326$. In both scenarios, the three-factor model cleanly separated the GEQ task, GEQ social, and leadership direction items into distinct factors. In the high cohesion scenario, we observed small-medium correlations between the complementarity factor and GEQ task, $r = 0.41$, and GEQ social, $r = 0.31$, and small correlations in the low cohesion scenario, $r = 0.26$ and $r = 0.28$, respectively.

Overall, these results suggest that the three-item measure of leadership direction has good psychometric properties of strong internal consistency, partial scalar invariance, sensitivity to depictions of high and low cohesion, and is both distinct from, and correlated with, task and social cohesion.

3.2.11 Resilience: Team Learning Orientation

For the team learning orientation (TLO) items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 28 and as many as 75 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of TLO did not fit the data well for high, $\chi^2(5) = 23.989$, $p < 0.001$, or low cohesion scenarios, $\chi^2(5) = 38.269$, $p < 0.001$. We did not get a model to fit the data with more than three items and selected the following items in our final model: "Mistakes are openly discussed in the team in order to learn from them" ($\lambda_{high} = 0.816$; $\lambda_{low} = 0.907$), "The team discusses performance constructively" ($\lambda_{high} = 0.881$; $\lambda_{low} = 0.956$), and "Team members learn and adapt with each other" ($\lambda_{high} = 0.702$; $\lambda_{low} = 0.830$). This model had good internal consistency for the high cohesion scenario, McDonald's $\omega = 0.84$, and excellent internal consistency for the low cohesion scenario, McDonald's $\omega = 0.93$.

Our configural-invariant model did not fit the data well, $\chi^2(8) = 19.921$, $p = 0.011$ These results indicate that our TLO measure may be too problematic to include in future measures.

A two-factor model produced poor model fit for both high, $\chi^2(19) = 97.225$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 135.641$, $p < 0.001$. The three-factor exploratory model did not fit well for the high cohesion scenario, $\chi^2(12) = 29.976$, $p = 0.003$, but it did fit for the low cohesion scenario, $\chi^2(12) = 8.097$, $p = 0.778$. In both scenarios, the three-factor model separated GEQ task, GEQ social, and TLO items into distinct factors. In the high cohesion scenario, we observed small correlations between the Perceived Efficacy for Collective Team Action (PECTA) factor and GEQ task, $r = 0.18$, and GEQ social, $r = 0.27$, and in the low cohesion scenario, $r = 0.29$ and $r = 0.36$, respectively.

Overall, these results suggest that if the TLO items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the attraction-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that

respondents may not interpret these statements about perceived efficacy in the same way in different contexts.

3.2.12 Resilience: Shared Language

For the shared language items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 40 and as many as 86 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of shared language fits the data well for the high cohesion scenario, $\chi^2(2) = 4.945$, $p = 0.103$, but not for the low cohesion scenarios, $\chi^2(2) = 6.994$, $p = 0.030$. After removing one item with very low factor loadings, we retained the following three items: "Both human and autonomous team members use common terms to understand one another" ($\lambda_{high} = 0.782$; $\lambda_{low} = 0.760$), "Team members use understandable communication patterns" ($\lambda_{high} = 0.815$; $\lambda_{low} = 0.836$), and "Team members are successful in understanding each other during missions" ($\lambda_{high} = 0.829$; $\lambda_{low} = 0.824$). This model had good internal consistency for both high and low cohesion scenarios, McDonald's $\omega = 0.85$ and McDonald's $\omega = 0.85$, respectively.

The configural invariant model fit the data well, $\chi^2(8) = 6.897$, $p = 0.548$, but the metric invariant model did not, $\chi^2(11) = 22.661$, $p = 0.020$. Interestingly, each item tested on its own was both metric and scalar invariant, but not when tested together. Even though the scalar invariant model did not fit when all items were tested simultaneously, $\chi^2(13) = 36.155$, $p = 0.001$, we retained it for the purposes of testing mean differences, rather than providing three separate mean comparisons depending on which item was modeled as invariant. According to this model, participants rated shared language in low cohesion scenarios 1.509 standard deviations lower than in high cohesion scenarios, $z = -12.826$, $p < 0.001$.

The two-factor exploratory model with shared language and GEQ items produced poor model fit for both high, $\chi^2(19) = 71.219$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 123.635$, $p < 0.001$. The three-factor exploratory model fit well for both the high, $\chi^2(12) = 18.247$, $p = 0.108$, and the low cohesion scenario, $\chi^2(12) = 7.735$, $p = 0.806$. In both scenarios, the three-factor model separated GEQ task, GEQ social, and shared language items into distinct factors, although in the high cohesion scenario, one task cohesion item had strong cross-loadings with social cohesion. In the high cohesion scenario, we observed small correlations between the shared language factor and GEQ task, $r = 0.29$, and GEQ social, $r = 0.19$, and

small-medium correlations in the low cohesion scenario, $r = 0.38$ and $r = 0.31$, respectively.

Overall, these results suggest that if the shared language items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the attraction-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. These items had good internal consistency and met criteria for partial scalar invariance, so they may be very helpful items in future measures.

3.2.13 Resilience: Team Flexibility

For the team flexibility items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, on these two items, 49 and 58 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

Our measure of team flexibility had only two indicators, which was not enough to fit a single-factor model on its own; to overcome this obstacle, we combined our team flexibility analyses with our shared language model. This model did not fit the data well for high, $\chi^2(4) = 16.790$, $p = 0.002$, or low cohesion scenarios, $\chi^2(2) = 16.217$, $p = 0.003$. There was no use in attempting to remove one of the two items, so rather than test for invariance, we examined whether these two items were subsumed by one of the factors in the GEQ.

Among the exploratory factor analyses, the only model that fit the data was the three-factor model of responses to the low cohesion scenario, $\chi^2(7) = 2.044$, $p = 0.957$, which clearly separated the team flexibility items from the two GEQ subfactors. These results indicate that the team flexibility subfactor does not have strong measurement properties. If it were to be included in future studies, it would need more than two indicators to provide the best opportunities for testing its properties.

3.2.14 Resilience: Perceived Efficacy of Collective Team Action

For the perceived efficacy of collective team action items, participants generally reported higher scores for the high cohesion versus the low cohesion scenario; still, as few as 50 and as many as 66 participants reported identical or lower scores on the high versus low cohesion scenario, indicating that they sincerely believed the situations were identical on those indicators or they were not responding sincerely. We did not remove any responses on this basis.

The single-factor model of our perceived efficacy items did not fit the data well for high, $\chi^2(9) = 19.235$, $p = 0.023$, or low cohesion scenarios, $\chi^2(9) = 68.66$, $p < 0.001$. We did not get a model to fit the data with more than three items and selected the following items in our final model: "The team is able to work together to accomplish the mission" ($\lambda_{high} = 0.869$; $\lambda_{low} = 0.879$), "The team can handle even the most difficult situations" ($\lambda_{high} = 0.839$; $\lambda_{low} = 0.951$), and "The team learns from challenges they face" ($\lambda_{high} = 0.887$; $\lambda_{low} = 0.850$). This model had excellent internal consistency for both high and low cohesion scenarios, McDonald's $\omega = 0.90$ and McDonald's $\omega = 0.92$, respectively.

Our configural-invariant model fit the data well, $\chi^2(8) = 6.445$, $p = 0.598$, but the metric-invariant model did not, $\chi^2(11) = 29.203$, $p = 0.002$. After testing items individually, we found one item to be metric invariant, "The team learns from challenges they face," but not scalar invariant. These results indicate that our PECTA measure may be too problematic to include in future measures.

We then conducted exploratory factor analysis of the PECTA along with the six-item GEQ. The two-factor model produced poor model fit for both high, $\chi^2(19) = 90.628$, $p < 0.001$, and low cohesion scenarios, $\chi^2(19) = 123.343$, $p < 0.001$. The three-factor exploratory model fit well for the high cohesion scenario, $\chi^2(12) = 10.714$, $p = 0.554$, and for the low cohesion scenario, $\chi^2(12) = 5.127$, $p = 0.954$. In both scenarios, the three-factor model separated GEQ task, GEQ social, and PECTA items into distinct factors, although in the high cohesion scenario, one task cohesion item had strong cross-loadings with social cohesion. In the high cohesion scenario, we observed small correlations between the PECTA factor and GEQ task, $r = 0.22$, and GEQ social, $r = 0.27$, and small-medium correlations in the low cohesion scenario, $r = 0.39$ and $r = 0.35$, respectively.

Overall, these results suggest that if the PECTA items are a good measure of cohesion and the GEQ-10 task and social cohesion items accurately measure their constructs, then the attraction-based cohesion items form a distinct factor that the GEQ-10 does not adequately capture. However, these items were not invariant, indicating that respondents may not interpret these statements about perceived efficacy in the same way in different contexts.

3.2.15 Results Summary

Taken together, the results of these analyses highlight several items with very good measurement properties, especially from the scales designed to assess perceived complementarity, morale, leadership direction, and perceived efficacy. However, some items with excellent properties still belong to measurement scales with apparent issues. Further, these analyses also revealed several problems with some

of the most commonly used scales for assessing team cohesion, such as the GEQ, measures of exclusivity, social cohesion, and belongingness, to name a few. Overall, these results will help guide recommendations for future measures of HAT cohesion.

4. Discussion

The primary goal for this research effort involved investigating the construct of team cohesion as it relates to HATs and developing a measure of cohesion that assesses this new and evolving construct. In this experiment, we pursued this goal by administering the proposed cohesion scale, identifying the items most relevant to the HATs depicted in the video clips, and removing the irrelevant items. By reducing the item pool in the current study, we set the stage for future validation of the cohesion construct, and continued item reduction, so this scale can be more easily used in future HAT experiments.

Using factor analysis methods, we were able to reduce our item pool substantially by removing items with low factor loadings and non-invariance. This reduction resulted in a self-report measurement scale that is concise, internally consistent, and invariant—the resulting items are both distinct from and correlated with the social and task cohesion subfactors of the GEQ (our criterion measure). Of the 82 items we used to assess 13 unique constructs, we found just nine items measuring only four constructs with these highly desirable measurement properties—four to measure perceived complementarity, three to measure morale, one to measure leadership direction, and one to measure perceived efficacy for collective team action (see Appendix A). We feel confident in recommending these high-quality items (see the *Items retained* column of Table 5). We cautiously recommend the use of the 28 items that partially met our prerequisites (see the *Items included* column of Table 5). While the remaining items did not meet our needs in this context, we recommend additional scrutiny of the other items from the remaining nine dimensions, given their specificity of our context and the utility these items had in other contexts.

Additionally, to establish credibility for our cohesion scale, it is important to first establish that the scale items are internally consistent and can, therefore, be trusted for use in measurements of other HATs, which was indeed found in our results. Results also showed that there was a strong correlation between items in the reduced scale's four subdimensions and their corresponding GEQ ratings, which indicates that our scale items for measurement of HATs align with the scale items that are accepted for measurement of cohesion in HATs. Finally, although not significantly different from one another, our preliminary results showed that

participants rated individual items of the scale much higher for the high cohesive HATs in comparison to the low cohesive HATs.

Several outcomes were found after analyzing the sample's demographic data. First, the overall immersive tendencies, involvement, and focus categories obtained via the ITQ were all below average, suggesting that participants were less liable to become absorbed in the video clips and more likely to be distracted from the video clips by environmental disturbances (Rózsa et al. 2022). Further, several relationships were found between affect and personality indicators and subjects' subsequent subjective cohesion ratings. Here, participants who self-reported high NA, obtained via the PANAS-X prior to the start of the experiment, also reported higher cohesion following "positive" cohesion clips. Additionally, correlations performed between demographic variables and subjective cohesion ratings for the high and low video clips revealed that the only significant relationships found were a negative relationship between *extraversion* and aggregate "positive" cohesion, a negative relationship between agreeableness and aggregate "positive" cohesion, and a negative relationship between agreeableness and total aggregate cohesion. These findings suggest that as participants' self-reported extraversion increases, their subjective ratings of "high cohesion" teams were reduced. Further, as participants' agreeableness increases, their ratings of cohesion in "high cohesion" teams specifically and all teams in general are reduced. This finding seems somewhat counterintuitive as anecdotally the construct of "agreeableness" may indicate more PA and generally good disposition; therefore, one may assume that subjective ratings of higher cohesion would also be reported for individuals with predispositions to this trait. A possible explanation for these findings may be that, in general, high cohesion teams exhibit more friendly, cooperative, and agreeable attitudes and behaviors (Lu 2015), which may be easier to observe and assess, compared to low cohesive teams whose behaviors may include more subtle nuances associated with poor performance. High cohesion facilitates teamwork behaviors that empower individuals and autonomous systems alike to continue to do their jobs effectively. Thus, participants may have had difficulty identifying slower, less cooperative behaviors and role uncertainty exhibited in the low cohesion vignettes, in comparison, which may reflect the difference in ratings here between high and low cohesive teams.

5. Limitations and Future Directions

This investigation of several established and novel HAT cohesion measures brought forth many valuable lessons that should be applied in subsequent work. First, we found that we were unable to conduct thorough analyses of the relationships between measurement scales because participants responded to

unique scenarios for virtually every measurement scale. Future studies could address this limitation by asking participants to respond to a single scenario when providing their ratings. Relatedly, our study design required participants to respond twice to each of the 82 non-GEQ items and 18 times to each of the GEQ items, possibly resulting in survey fatigue and thereby compromising the sincerity of participants' responses. Several items with strikingly similar content performed quite differently on our various metrics, indicating that people may not have been attending as closely to the differences and similarities in the items' content. And while most participants reported higher scores for the high cohesion scenarios, nearly one quarter of all participants rated the low cohesion scenarios higher than, or equal to, the high cohesion scenarios, providing further evidence that participants were not attending to the stimuli and/or item content.

We expect that asking participants to respond to fewer items may help alleviate these issues in the future. In addition to reducing the size and frequency of questionnaires, we also recommend that future investigations of these measures include many more respondents, perhaps more than 500 (Jiang et al. 2016). Doing so would enable us to understand much more about these scales and items, as the larger sample size will open the doors to testing item response theory models. For instance, the graded response model (Samejima 1997) allows for testing item information on ordinal response scales, which is more appropriate for these data than treating the ordinal responses as continuous. Finally, we found out during analyses that our selected criterion measure of cohesion, the GEQ-10 (Carless and De Paola 2000), was poorly represented by the theoretical three-factor model representing social cohesion, task cohesion, and individual attraction to the group. We discuss these issues further in the remainder of this section.

Concurrently with Carless and De Paola's publication of the 10-item version of the GEQ, which was shortened from the 18-item scale developed by Carron et al. (1985), Carron and Brawley (2000) challenged the use of the GEQ beyond the sport team context. In their critique, Carron and Brawley noted several instances in which scholars sought to use the GEQ to measure team cohesion in various groups, such as musicians in a band, dyadic work teams, and even high school athletes in various team sports. In each of these cases, they noted that researchers could not recreate the factor structure posited by Carron et al. (1985), thus calling into question both the theoretical structure of cohesion and, importantly, the measurement items used to assess that structure. Considering that we encountered very similar issues in our analyses of the GEQ-10, we would recommend identifying a different criterion measure of cohesion, perhaps by asking participants to respond to a few items that explicitly call their attention to team cohesion. Interestingly, none of the GEQ-10 items mention the word *cohesion* at all, and only one item from our other scales

used the word cohesion. Perhaps using more direct indicators will give us clearer information about whether, and how, responses to these other scales relate to people's understanding of what dimensions contribute to a cohesive team.

6. Conclusions and Path Forward

By using factor analysis methods, we were able to identify nine items across four dimensions of cohesion: perceived complementarity, morale, leadership direction, and perceived efficacy that exhibited good or very good psychometric properties. In addition to identifying excellent indicators, we also identified several areas to improve in future studies that will help us solidify our understanding and recommend a more comprehensive scale of HAT cohesion. Future efforts involve conducting further experiments with the parsed down item pool to obtain a larger sample size, as well as utilizing a different criterion measure due to the problematic nature of the GEQ-10 that was currently used. Once this is accomplished and we have obtained enough information to produce a final measurement scale, we will be able to provide other researchers and practitioners with clearer guidance on how to administer the scale and calculate respondents' scores along the different dimensions. Overall, these results are encouraging as we were able to identify several areas where we need to improve our understanding of HATs, both methodologically and theoretically. In addition to the methodological improvements proposed previously, we found that some of our best cohesion measures in this study are often overlooked in other investigations—specifically perceived complementarity, leadership direction, and perceived efficacy.

7. References

Abrams AM, Rosenthal-von der Pütten AM. I–C–E framework: concepts for group dynamics research in human-robot interaction. International Journal of Social Robotics. 2020;1–17.

Ahronson A, Cameron J. The nature and consequences of group cohesion in a military sample. Military Psychology. 2007;19:9–25. 10.1080/08995600701323277.

Anthoine E, Moret L, Regnault A, Sébille V, Hardouin JB. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. Health Qual Life Outcomes 2014;12:176.

Bandura A. Self efficacy: the exercise of control. W. H. Freeman and Company; 1997.

Barnes MJ, Evans AW III. Soldier-Robot teams in future battlefields: an overview. In: Barnes M; Jentsch F, editors. Human-Robot Interactions In Future Military Operations. Ashgate; 2010. p. 9–29.

Beal DJ, Cohen RR, Burke MJ, McLendon CL. Cohesion and performance in groups: a meta-analytic clarification of construct relations. Journal of Applied Psychology. 2003;88(6):989–1004. https://doi.org/10.1037/0021-9010.88.6.989.

Berg S, Neubauer C, Robison C, Kroninger C, Schaefer KE, Krausman A. Exploring resilience and cohesion in human autonomy teams: models and measurement. Proceedings of the 12th International Conference on Applied Human Factors and Ergonomics; 2021; New York, NY.

Boateng GO, Neilands T, Frongillo E, Melgar-Quiñonez H, Young S. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Frontiers in Public Health. 2018;6.

Bollen KA, Hoyle RH. Perceived cohesion: a conceptual and empirical examination. Social Forces. 1990;69(2):479–504. https://doi.org/10.2307/2579670.

Bowers C, Kreutzer C, Cannon-Bowers J, Lamb J. Team resilience as a second-order emergent state: a theoretical model and research directions. Frontiers in Psychology. 2017;8. https://doi.org/10.3389/fpsyg.2017.01360.

Carless SA, De Paola C. The measurement of cohesion in work teams. Small Group Research. 2000;31(1):71–88.

Carron AV, Brawley LR. Cohesion. Small Group Research. 2000;31(1):89–106. https://doi.org/10.1177/104649640003100105.

Carron AV, Spink KS. Team building in an exercise setting. The Sport Psychologist. 1993;7(1):8–18.

Carron AV, Widmeyer WN, Brawley LR. The development of an instrument to assess cohesion in sport teams: the Group Environment Questionnaire. Journal of Sport Psychology. 1985;7:244–266.

Cato CR, Blue SN, Boyle B. Conceptualizing risk and unit resilience in a military context. In: Trump BD, Florin M-V, Linkov I, editors. IRGC resource guide on resilience. EPFL International Risk Governance Center; 2018. (vol. 2): Domains of resilience for complex interconnected systems. irgc.org.

Chen JYC, Barnes MJ. Human-agent teaming for multirobot control: A review of human factors issues. IEEE Transactions on Human-Machine Systems. 2014;44(1):13–29. https://doi.org/10.1109/THMS.2013.2293535.

Chen JY, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science. 2018;19(3):259–282.

Chiocchio F, Essiembre H. Cohesion and performance: a meta-analytic review of disparities between project teams, production teams, and service teams. Small Group Research. 2009;40(4):382–420.

Comrey AL. Factor-analytic methods of scale development in personality and clinical psychology. Am Psychol Assoc. 1988;56:754–61.

Craig TY, Kelly JR. Group cohesiveness and creative performance. Group Dynamics: Theory, Research, and Practice. 1999;3(4): 243–258. doi:10.1037/1089-2699.3.4.243.

Crisp RJ, Stone CH, Hall NR. Recategorization and subgroup identification: Predicting and preventing threats from common ingroups. Personality and Social Psychology Bulletin. 2006;32:230–243.

Demir M, Likens A, Cooke NJ, Amazeen P, McNeese NJ. Team coordination and effectiveness in human-autonomy teaming. IEEE Transactions on Human-Machine Systems. 2019;49(2):150–159.

Dion KL. Group cohesion: from "field of forces" to multidimensional construct. Group Dynamics: Theory, Research, and Practice. 2000;4(1):7–26. https://doi.org/10.1037/1089-2699.4.1.7.

Donnellan M, Oswald F, Baird B, Lucas R. The Mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychological Assessment. 2006;18:192–203. 10.1037/1040-3590.18.2.192.

Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology. 2014;105(3):399–412.

Edmondson AC. Psychological safety and learning behavior in work teams. Adm Sci Q. 1999;44(2):350–83.

Evans NJ, Jarvis PA. Group cohesion: a review and reevaluation. Small Group Behavior. 1980;11(4):359–370. doi:10.1177/104649648001100401.

Forsyth DR. Group dynamics. 3rd ed. Brooks/Cole; 1999.

Friedkin N. Social cohesion. Annual Review of Sociology. 2004;30:409–425. 10.1146/annurev.soc.30.012703.110625.

Gaertner SL, Dovidio JF. A common ingroup identity: a categorization-based approach for reducing intergroup bias. Handbook of Prejudice. In: Nelson T, editor. Taylor and Francis; 2009. p. 489–506.

Gittell JH, Cameron K, Lim S, Rivas V. Relationships, layoffs, and organizational resilience airline industry responses to September 11. J Appl Behav Sci. 2006;42:300–329. doi: 10.1177/0021886306286466.

Goldberg LR. International personality item pool: a scientific collaboratory for the development of advanced measures of personality traits and other individual differences; 1999 [accessed 2007 Dec 6]. http://ipip.ori.org.

Griffin A. The effect of project and process characteristics on product development cycle time. Journal of Marketing Research. 1997;34(1):24–35.

Griffith J. Measurement of group cohesion in US Army units. Basic and Applied SocialPpsychology. 1988;9(2):149–171.

Griffith J, Vaitkus M. Relating cohesion to stress, strain, disintegration, and performance: an organizing framework. Military Psychology. 1999;11(1):27–55.

Grossman R. How do teams become cohesive? [Dissertation]. University of Central Florida; 2014 unpublished.

Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. Am Psychol Assoc. 1988;103:265–75. 10.1037/0033-2909.103.2.265.

Hogg MA. The social psychology of group cohesiveness: from attraction to social identity. Harvester Wheatsheaf; 1992.

Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. Frontiers in Psychology; 2016;7:109.

Lakhmani SG, Neubauer C, Krausman A, Fitzhugh SM, Berg SK, Wright JL, Schaefer KE. Cohesion in human–autonomy teams: an approach for future research. Theoretical Issues in Ergonomics Science. 2022;1–38.

Langfred CW. Is group cohesiveness a double-edged sword? An investigation of the effects of cohesiveness on performance. Small Group Research. 1998;29(1):124–143.

Lawshe CH. A quantitative approach to content validity. Personnel Psychology. 1975;28(4):563–575.

Looije R, Neerincx MA, Cnossen F. Persuasive robotic assistant for health self-management of older adults: design and evaluation of social behaviors. International Journal of Human-Computer Studies. 2010;68(6):386–397.

Lott AJ, Lott BE. Group cohesiveness as interpersonal attraction: a review of relationships with antecedent and consequent variables. Psychological Bulletin. 1965;64(4):259.

Lu L. Building trust and cohesion in virtual teams: The developmental approach. Journal of organizational effectiveness: People and performance. 2015;2(1): 55–72.

MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. Psychol Methods. 1999;4:84–99. 10.1037/1082-989X.4.1.84.

Mathieu JE, Kukenberger MR, D'Innocenzo L, Reilly G. Modeling reciprocal team cohesion-performance relationships, as impacted by shared leadership and members' competence. Journal of Applied Psychology. 2015;100(3):713–734. https://doi.org/10.1037/a0038898.

Morgado FFR, Meireles JFF, Neves CM, Amaral AC, Ferreira ME. Scale development: ten main limitations and recommendations to improve future research practices. Psicol Refl Crít. 2018;30(3). https://doi.org/10.1186/s41155-016-0057-1.

Morgan PB, Fletcher D, Sarkar M. Defining and characterizing team resilience in elite sport. Psychology of Sport and Exercise. 2013;14(4):549–559.

Morrow PB, Fiore SM. Supporting human-robot teams in social dynamicism: an overview of the metaphoric inference framework. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2012 Sep. Los Angeles, CA. SAGE Publications. Vol. 56, No. 1, p. 1718–1722.

Muchinsky PM, Monahan CJ. What is person environment congruence? Supplementary versus complementary models of fit. Journal of Vocational Behavior. 1987;31(3):268–277. doi:10.1016/0001-8791(87)90043-1.

Mullen B, Copper C. The relation between cohesiveness and productivity: An integration. Psychological Bulletin. 1994;115:210–227.

Mudrack PE. Group cohesiveness and productivity: A closer look. Human Relations. 1989;42(9):771–785.

Neubauer C, Woolley J, Khooshabeh P, Scherer S. Getting to know you: a multimodal investigation of team behavior and resilience to stress. In Proceedings of the 18th ACM International Conference on Multimodal Interaction; 2016 Oct. p. 193–200.

Nomura T, Suzuki T, Kanda T, Kato K. Altered attitudes of people toward robots: Investigation through the negative attitudes toward robots scale. AAAI Workshop; 2006. Technical Report.

Norris FH, Stevens SP, Pfefferbaum B, Wyche KF, Pfefferbaum RL. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. Am J Community Psychol. 2008;41,127–150. doi: 1007/s10464-007-9156-6.

Oosterhof A, Van der Vegt GS, Van de Vliert E, Sanders K. Valuing skill differences: Perceived skill complementarity and dyadic helping behavior in teams. Group and Organization Management, 2009;34(5):536–562.

O'Reilly CA 3rd, Caldwell DF, Barnett WP. Work group demography, social integration, and turnover. Administrative science quarterly. 1989: 21–37.

Osborne JW, Costello AB. Sample size and subject to item ratio in principal components analysis. Pract Assess Res Eval. 2004;99:1–15. http://pareonline.net/htm/v9n11.htm.

Phillips E, Ososky S, Grove J, and Jentsch F. From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting; Los Angeles, CA. SAGE Publications. 2011;55(1):1491–1495.

Piasentin KA, Chapman DS.. Perceived similarity and complementarity as predictors of subjective person-organization fit. Journal of Occupational and Organizational Psychology. 2007;80(2):341–354.

Putnam RD. Tuning in, tuning out: the strange disappearance of social capital in America. PS: Political Science and Politics, vol. 28, no. 4, 1995.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2020. https://www.R-project.org.

Revelle W. Psych: Procedures for personality and psychological research. Northwestern University; 2022. https://CRAN.R-project.org/package=psych.

Rosseel R. lavaan: an R package for structural equation modeling. Journal of Statistical Software. 2012;48(2):1–36. https://doi.org/10.18637/jss.v048.i02.

Rózsa S, Hargitai R, Láng A, Osváth A, Hupuczi E, Tamás I, Kállai J. Measuring immersion, involvement, and attention focusing tendencies in the mediated environment: the applicability of the immersive tendencies questionnaire. Frontiers in Psychology. 2022;13.

Russell S, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Proceedings of Prentice Hall series in artificial intelligence. Prentice Hall; 2010.

Salas E, Dickinson TL, Converse SA, Tannenbaum SI. Toward an understanding of team performance and training. In: Swezey RW, Salas E, editors. Teams: their training and performance. Ablex Publishing; 1992. p. 3–29.

Salas E, Grossman R, Hughes A, Coultas CW. Measuring team cohesion. Human Factors: The Journal of Human Factors and Ergonomics Society. 2015;57:365–374.

Salas E, Rico R, Passmore J, Vessey WB, Landon LB. The Wiley Blackwell handbook of the psychology of team working and collaborative processes. Wiley; 2017.

Salas E, Sims DE, Burke CS. Is there a “Big Five” in teamwork? Small Group Research. 2005;36(5):555–599

Samejima F. Graded response model. In: Handbook of modern item response theory. Springer; 1997. p. 85–100.

Schaefer KE, Oh J, Aksaray D, Barber D. Integrating context into artificial intelligence: research from the robotics collaborative technology alliance. AI Magazine. 2019;40(3):28–40. https://doi.org/10.1609/aimag.v40i3.2865.

Schaefer KE, Sanders TL, Yordon RE, Billings DR, Hancock PA. Classification of robot form: factors predicting perceived trustworthiness. In: Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting; 2012; Santa Monica, CA. Human Factors and Ergonomics Society. p. 1548–1552.

Schaefer KE, Straub ER, Chen JYC, Putney J, Evans AW. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. Cognitive Systems Research. 2017;46:26–39.

Sharma S, Sharma SK. Team resilience: scale development and validation. Vision. 2016;20(1):37–53.

Siebold G. The evolution of the measurement of cohesion. Military Psychology. 1999;11:5–26. 10.1207/s15327876mp1101_2.

Siebold G, Kelly DR. The impact of cohesion on platoon performance at the joint readiness training center. US Army Research Institute for the Behavioral and Social Sciences; 1988. Technical Report, Vol. 812.

Siebold GL. Military group cohesion. Military life: The psychology of serving in peace and combat. 2006;1:185–201.

Smith KG, Smith KA, Olian JD, Sims HP, O’Bannon DP, Scully JA. Top management team demography and process: the role of social integration and communication. Administrative Science Quarterly. 1994;39(3):412. https://doi.org/10.2307/2393297.

Stuster J. Bold endeavors: lessons from polar and space exploration. Naval Institute Press; 1996.

Sycara K, Sukthankar G. Literature review of teamwork models. Carnegie Mellon University; 2006. Report No.: CMU-RITR-06-50.

Taylor GS, Barnett JS. Training capabilities of wearable and desktop simulator interfaces. Army Research Institute for the Behavioral and Social Sciences; 2011.

Taylor G, Singer MJ, Jerome CJ. Development and evaluation of the game-based performance assessment battery (GamePAB) and game experience measure (GEM). Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2009 Oct. Los Angeles, CA: SAGE Publications. Vol. 53, No. 27, p. 2014–2018.

Tsui KM, Desai M, Yanco H, Cramer H, Kemper N. Using the negative attitude toward robots scale with telepresence robots. Proceedings of the 10th Performance Metrics for Intelligent Systems Workshop on PerMIS 2010. 2010. p. 243.

Walliser JC, de Visser EJ, Wiese E, Shaw TH. Team structure and team building improve human–machine teaming with autonomous agents. Journal of Cognitive Engineering and Decision Making. 2019;13(4):258–278.

Watson D, Clark LA. The PANAS-X: manual for the positive and negative affect schedule-expanded form. Iowa Research Online; 1999 [accessed 2015 Sep 29]. http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs.

Williams J, Brown JM, Bray RM, Anderson Goodell EM, Olmsted KR, Adler AB. Unit cohesion, resilience, and mental health of soldiers. In: Basic Combat Training. Military Psychology. 2016;28(4):241–250. https://doi.org/10.1037/mil0000120

Witmer BG, Singer MJ. Measuring presence in virtual environments: a presence questionnaire. Presence. 1998;7(3):225–240.

Zaccaro SJ. Nonequivalent associations between forms of cohesiveness and group-related outcomes: evidence for multidimensionality. The Journal of Social Psychology. 1991;131(3):387–399.

Appendix A. Original Item Pool with Item Retention Recommendations

Function-based Cohesion

a. Task Cohesion

2) +Our team was committed to working together
3) Our team was engaged in the task
4) +Our team accomplished assigned tasks to the best of their ability
5) +Our team was united in working toward task goals
6) Our team members had conflicting task goals (R)
7) All team members are dedicated to doing their jobs well
8) All team members helped each other to get the job done
9) All team members worked hard to accomplish the task
10) All team members pulled together to share the load while performing the task
11) Our team was successful because team members worked together

Structural Cohesion

a. Exclusivity

12) All team members are expected to share responsibility for poor performance
13) All team members share responsibility for loss or poor performance
14) Team members acknowledge successful team performance
15) Members of my team will readily defend decisions made by autonomous teammates
16) +This team has clearly established norms for working together
17) +Most of the team members fit what I feel to be the idea of a good teammate
18) +Team members embody the ideal for a good team member
19) Team member behavior standards are vague and unclear (R)
20) It is clear what is and what is not acceptable member behavior in the team

b. Individual attraction to team

21) +I would not miss the members of this team (R)
22) +If given the chance, I would choose to leave my team and join another (R)

23) +If given the chance, I would remain a member of this team

24) If given the chance, my teammates would have me leave the team and join another (R)

c. Leadership Direction (vertical cohesion)

25) Our team members respect the leaders in this team

26) When an individual in this team needs help, their leaders acknowledge those needs and their importance

27) *Leaders understand the capabilities of this team

28) Leaders keep themselves informed about the progress team members are making during the task

29) +The leaders support team members during task performance

30) +The leaders work along with their team members

Interpersonal Cohesion

a. Team Pride

31) +Membership in this team is an accomplishment worthy of pride

32) +Team members acknowledge the importance of the team's mission

33) +Both human and autonomous team members know their importance in accomplishing task goals

b. Social Cohesion

34) I believe other team members liked me

35) Both human and autonomous team members listened to what I had to say

36) Observing my teammates successful behaviors and actions helped me to stay on task

37) Members of our team would rather work alone than as a team (R)

38) I would like to work with the same team members on a similar task

39) +I liked the team I was in

40) +I enjoyed interacting with this team

41) There was a feeling of unity and cohesion in the team

42) The members of my team are close to one another

43) Both human and autonomous members of this team get along well

44) I generally do not get along with the other members of my team (R)

45) Human and autonomous team members care about what happens to each other

46) Members like being in this team

47) +Members in this team respect one another

48) Members in this team like one another

a. Belongingness

49) +I do not feel a sense of belonging to this team (R)

50) +I feel that I am a member of this team

51) +I feel connected to my human and non-human team members

b. Morale

52) Working with my team makes me feel good

53) *My personal feelings about working with both human and autonomous team members were negative (R)

54) *I am unhappy with the level of shared awareness within my team (R)

55) *I do not like how human and autonomous team members work together (R)

Perceived complementarity

56) *Individual members of the team are important because they offer skills and abilities that work well together

57) I feel that I am important to this team because I have skills and abilities that work well with my teammates

58) *My teammates rely on me because I have skills that they do not have

59) When key decisions are made, my teammates consult me because I have a different perspective than they do

60) My knowledge, skills and/or abilities offer something that other members in this team do not have

61) I feel that I am a unique piece of the puzzle that makes this team successful

62) *The skills of the autonomous teammate(s) complement me in things I am not good at

63) *The other members and I compensate for each other's weaknesses

64) The other members and I can do more together than we can separately

65) The team's members are too dissimilar to work together well (R)

66) The team is diverse with each team member bringing a different perspective in a way that bolsters team success

Resilience

a. **Mastery Approaches:**
b. **Team Learning Orientation**

66) +Mistakes are openly discussed in the team in order to learn from them

67) +The team discusses performance constructively

68) The same mistakes are made over and over in the team (R)

69) Team members are encouraged to ask questions to gain a deeper understanding of their goals

70) +Team members learn and adapt with each other

a. **Team Flexibility**

71) ΔTeam members are capable of adjusting their approach(es) to overcome obstacles

72) ΔTeam members are successful in handling stressful tasks and missions

a. **Social Capital:**
b. **Shared Language**

73) +Both human and autonomous team members use common terms to understand one another

74) +Team members use understandable communication patterns

75) +Team members are successful in understanding each other during missions

76) Team members did not communicate well during the task (R)

a. **Collective Efficacy:**
b. **Perceived Efficacy for Collective Team Action**

77) When a team member has a problem, the rest of the team is able to assist them

78) +The team is able to work together to accomplish the mission

79) +The team can handle even the most difficult situations

80) The team is capable of solving problems together

81) *The team learns from challenges they face

82) I can count on my human and autonomous team members

Note: (R) denotes reverse scored item; * denotes items with excellent psychometric properties that we recommend including in future measures (i.e., 9 total items); plus (+) denotes items with some promising measurement qualities that we recommend researching further (i.e., 28 total items); Δ denotes items that could not be adequately tested due to the small number of items in that subdimension, but may warrant further investigation (i.e., 2 total items).

Appendix B. Complete Video Game Experience Data

Participants were asked to report, on average, how many hours per week they currently play video games to determine if our sample was composed of a particularly high or low number of experienced video gamers. If so, this may impact the responses to our video vignette methodology. The following figure illustrates that 83% of our participant sample self-reported that they play video games less than 10 h a week (Fig. B-1). Furthermore, almost half of our sample (i.e., *N*= 120) self-reported that they most they have played is less than 10 h a week, suggesting that these participants could be considered gaming novices or even non-players.

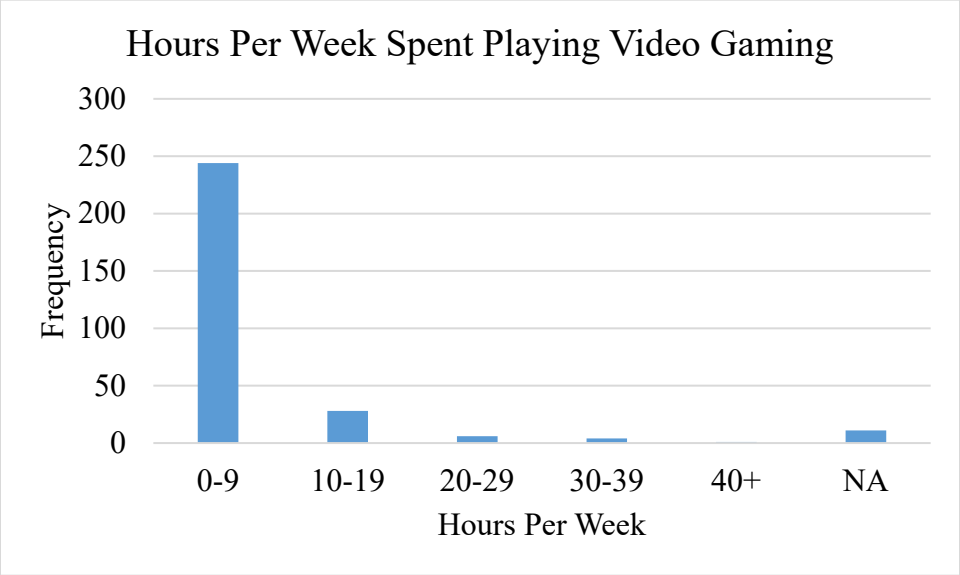

Fig. B-1 Frequency of hours per week spent playing video games

Other subdimensions of the video game experience questionnaire (e.g., graphical adventure games, puzzle games, turn-based strategy games) are omitted from this report as we did not believe they would influence subjective responses to our vignettes or the video clips.

Appendix C. Complete Immersive Tendencies Overall Statistics and Factors Data

Prior to the experiment, participants also completed the immersive tendencies questionnaire (ITQ), which yields an overall Immersive Tendencies score as well as scores for three subcategories of immersion, which included involvement, focus, and gaming. Figure C-1 illustrates our sample population's averaged score for overall immersive tendencies and for each of the three subcategories of the ITQ. Overall immersive tendencies, involvement, focus, and gaming are all below average, as indicated in Fig. B-1.

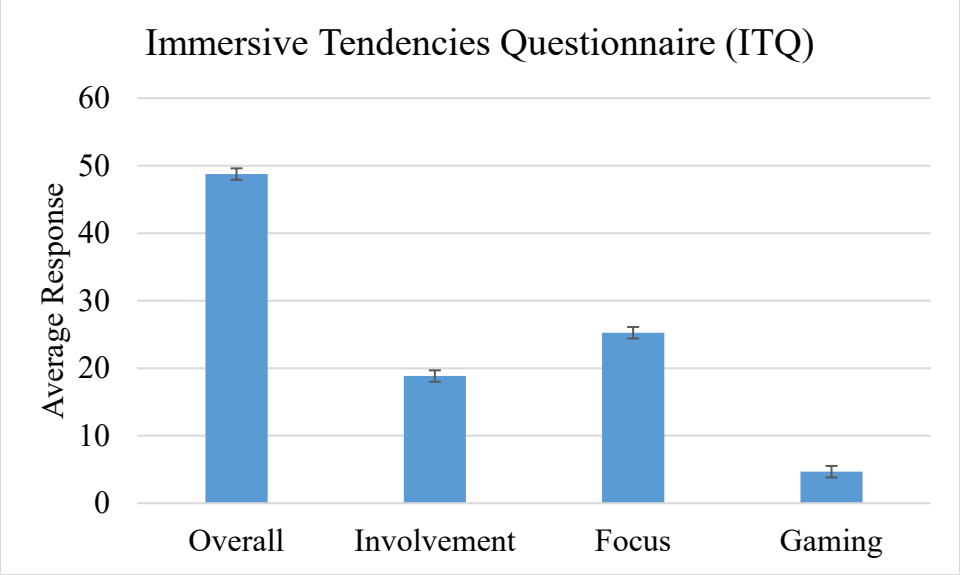

Fig. C-1 Participants' average overall immersive tendencies (i.e., overall) scores, as well as average scores on the Involvement (Involvement), Attention Focus (Focus), and Commitment to Games (Gaming) subscales. High scores indicate a greater tendency towards immersion, while lower scores indicate a lesser tendency towards immersion. Involvement scores can range from 7 to 49. Focus scores can range from 1 to 49. Gaming scores can range from 2 to 14. Overall scores can range from 16 to 112.

Appendix D. Complete Negative Attitudes Toward Robots (NARS) Overall Statistics and Subdimensions Data

To gauge any pre-existing attitudes towards robots, participants were instructed to complete the Negative Attitude Towards Robots scale prior to the experiment. The NARS provides information on three subdimensions: negative attitudes towards situations of interaction with robots (i.e., interaction), negative attitude towards social influence of robots (i.e., social influence), and negative attitude toward emotions in interaction with robots (i.e., emotion). As seen in Fig. D-1, responses across participants were averaged and revealed that participants reported social and emotion scores greater than average.

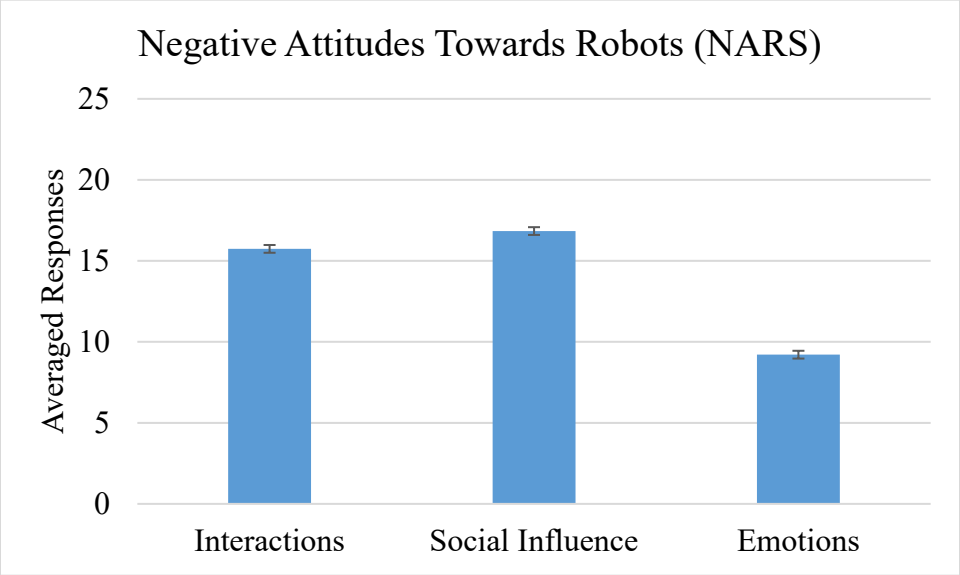

Fig. D-1 Averaged responses for the three NARS subscales. The minimum and maximum scores are 6 and 30 for the interaction subscale, 5 and 25 for social influence subscale, and 3 and 15 for the emotion subscale, respectively.

Appendix E. Item Pool for the GEQ-10

1) Our team is united in trying to reach its goals for performance
2) I'm unhappy with my team's level of commitment to the task (R)
3) Our team members have conflicting aspirations for the team's performance (R)
4) This team does not give me enough opportunities to improve my personal performance (R)
5) Our team would like to spend time together outside of work hours
6) Members of our team do not stick together outside of work time (R)
7) Our team members rarely party together (R)
8) Members of our team would rather go out on their own than get together as a team (R)
9) For me this team is one of the most important social groups to which I belong
10) Some of my best friends are in this team

List of Symbols, Abbreviations, and Acronyms

ARL	Army Research Laboratory
CFA	confirmatory factor analysis
DEVCOM	US Army Combat Capabilities Development Command
EFA	exploratory factor analysis
GEQ	Group Environment Scale
HAT	human-autonomy team
IA	intelligent agent
ITQ	Immersive Tendencies Questionnaire
NA	Negative Affect
NARS	Negative Attitudes toward Robots Scale
PA	Positive Affect
PANAS-X	Positive and Negative Affect Schedule Extended
PECTA	Perceived Efficacy for Collective Team Action
SME	subject matter expert
TLO	Team Learning Orientation
USMA	US Military Academy
VGE	Video Game Experience

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB
5 (PDF)	DEVCOM ARL FCDD RLA FA C NEUBAUER S LAKHMANI D FORSTER A KRAUSMAN S M FITZHUGH