Automated observation of competency-related behavior in serious gaming

Master Thesis

Author:

Martijn van den Berg (s1613367)

University supervisors

Dr. Khiet Truong

Prof. dr. Celeste Wilderom

Accenture supervisors

Rend Barnhoorn

Dr. Ivo Wenzler

Rutger Deenen

Table of Contents

Automated observation of competency-related behavior in serious gaming

Martijn van den Berg

University of Twente, the Netherlands

This exploratory study uses a cooperative digital game to explore possibilities for automatic observation of competency-related behavior in listening, verbal communication, taking initiative and decision making. During game play, audio and video of participants is recorded. Audio and video is analyzed using vocal and facial expression analysis. Demonstration of competency-related behavior is verified using a self-report and peer survey, completed by participants after game play. Self-report and peer surveys are correlated with vocal, facial expression and game data to determine to what extent competency-related behaviors can be predicted using computerized observations. Results show that, although behavior can to a large extent be predicted using game data, vocal and facial expression analysis, the predictors of these behaviors do not logically explain the behavior predicted. Several propositions are developed to help guide future research.

Key words: competency-related behavior, selection assessment, serious gaming, social signal processing, HR technology

Introduction

Personnel selection aims to find a candidate for a specific work environment. Job-environment fit consists of person-job fit and person-organization fit (Segikuchi, 2004). The most common methods used in selection are job interviews and behavioral assessments in assessment centers. In general, assessment centers are more valid at predicting job environment fit, because this method uses behavior instead of past achievements, thereby predicting future performance rather than current potential (Bartram, 2012).

Several researchers have focused on improving the predictive ability of selection assessments. However, several limitations of selection assessment still apply. Some limitations to current selection assessments include 1. behavioral assessments use multiple observations for observing behavior, which might create differences in interpretation of behavior (Green et al., 2011), 2. behavioral assessments use a combination of exercises rather than actual workplace performance, which might create differences between predicted workplace performance and actual workplace performance, 3. behavioral assessment reliability is increased by using multiple observers, but this costs considerably more (König et al., 2010, Kaslow et al., 2007), 4. behavioral assessment participants are aware of being observed, which might lead participants to show socially desirable behavior (Bangerter et al., 2012).

Digital serious games have been used as an alternative to traditional assessments (Chin et al., 2009). Serious games are games used for other purposes than mere entertainment (Susi et al., 2007). By creating memories of the future, games are able to simulate elements of prospective work environments, allowing behavior of participants to be assessed (Wenzler, 2008). The development of serious games over the last years has led to research to serious games increasingly being used as a selection assessment tool (Fetzer, 2015).

The combination of serious games with automated measurement methods can provide a more accurate alternative to traditional assessment because 1. (digital) serious games can make use of computerized observations, providing a more reliable way of collecting data (Tippins, 2011), 2. Serious games are able to simulate (elements of) a real life work environment, allowing participants to experience work situations before participating in a workplace environment (Wenzler, 2008), 3. Using computerized observation methods within serious games can possibly allow more reliable assessments at little additional costs (Fetzer, 2015) and 4. Social desirability might be reduced because serious games allow a high level of engagement, making participants less aware of behavior during selection assessments (Shute & Kim, 2013).

Most measurement methods within serious games use a form of game metrics to measure competency. These metrics are related to outcomes, either efficiency or effectiveness in various situations (Mayer et al., 2013). However, effective assessment requires not only outcomes to be measured, but also the process through which outcomes are achieved (Shute & Kim, 2013). Measuring competency-related behavior is more effective when more direct methods of measuring behavior are used (Belotti et al., 2013, Mayer et al., 2013). Therefore, this study aims to find how competency-related behavior during a serious game can be automatically identified and measured to document and visualize the presence or absence of participant competency-related behavior.

This study documents the possibilities of game data, vocal-emotion and facial expression analysis for directly measuring competency-related behavior in serious gaming. In this way this study contributes towards the field of personnel selection by exploring possibilities for automatization of observation. This contribution can be extended towards to field of assessment in general, because automatized observations methods are also useful for other types of assessments. In addition, finding automatized methods of measuring behavior will also contribute towards the field of serious gaming, expanding opportunities for using serious games in personnel selection. This research is explorative, because little research into automated measurement of competency-related behavior is currently available (Fetzer 2015).

First, an overview of the related works is given, describing issues with current selection methods as well as the current state of serious gaming in assessment and previous attempts at computerized measurement of behavior. Next, an overview of methods is given, describing research design, sample, instruments, measurements and procedures used. Next, an overview of results is given, discussing descriptive results of automated measurement methods, as well as models for predicting competency-related behavior using automated observation methods. Last, results are discussed to reach future propositions for the possibilities of observing competency-related behavior using automated methods, leading to a general conclusion.

Context

Current selection methods and limitations

Personnel selection aims at using a reliable and valid way in which actual performance in a workplace environment can be predicted. This is commonly referred to as person-environment fit (Segikuchi, 2004). Person environment fit consists of the ability of an individual to work in a specific function, as well as within a specific organization culture. Most often person-environment fit is predicted either as testing a set of knowledge, skills, abilities and other, also referred to as the KSAO model (Cheney et al., 1990) or sets of behavior which represent workplace performance (Bartram, 2012). Using the KSAO model predicts competence, whereas approaches using behavior for predicting person-environment fit use competency. Using competency to predict person-environment fit is more accurate, because competency reflects future performance rather than current achievements (Bartram, 2012).

There are two main differences between competence and competency. The first difference is that competency refers to developmental abilities rather than just dimensions of performance (Kaslow et al., 2007) and is therefore more forward looking, whereas competence is based on previous abilities required to perform a job (Bartram, 2012). Second, competence refers to the potential of an individual to perform in a workplace environment, whereas competency consists of actual observable behaviors shown by application of competence in a workplace environment (Figure 1). Predicting actual person-environment fit is often difficult, because this requires job tasks to be broken down into either KSAO’s to determine competence or recognizable behaviors to determine competency, and allow each of these to be tested. Breaking down job

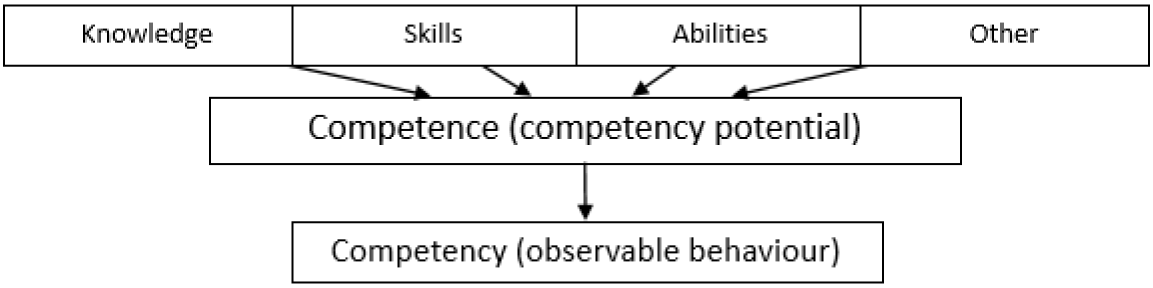

Knowledge
Skills
Abilities
Other
Competence (competency potential)
Competency (observable behaviour)

FIGURE 1: RELATIONSHIP BETWEEN COMPETENCE AND COMPETENCY

requirements into either competences or competencies is also known as job analysis or competency modelling respectively (Ryan & Ployhart, 2014, Delamare la Deist & Winterton, 2005).

Common methods used for predicting workplace performance are job interviews and assessments in assessment centers. Job interviews aim at testing KSAO's trough interview questions and predict behavior trough situational judgement tests (Christian et al., 2010). Assessment centers use a combination of interviews and specific exercises. The choice on either of these methods is often a trade-off between validity and financial resources, as assessments are often more valid in predicting environment behavior yet are more cost extensive (König et al., 2010, Kaslow et al., 2007). The main reason for the higher predictive validity of assessment centers is the ability to test competency in workplace simulations instead of questions representing future use of competency or past competence (Bartram, 2005).

Reliability of test results in an assessment center requires multiple assessors to observe participants during the execution of specific exercises across a multitude of dimensions (Putka & Hoffman, 2013). Each observer is required to have a high degree of psychometric expertise in order to accurately assess the behavior of participants (Chin et al., 2009). In addition, objective measurement of soft competencies requires consistency among assessors, which is often difficult due to the complex nature of these competencies (Green et al., 2011). Therefore, having a reliable assessment requires a number of observers, which is often costly.

Reliability within assessment centers becomes more complex when taking into consideration the reliability of results between assessors (Putka & Hoffman, 2013). Although metrics can be established to score participants on various dimensions, the interpretation of these results might differ between the assessment centers. Accurately establishing a unified scoring system requires quantitative measures to be set up which can be used across assessment centers. However, differences in culture prevent such measures from becoming accurate (Ryan & McFarland, 1999). Assessing across cultural borders requires different standards of behavior assessment to be established. In addition, the role of organizational culture is increasingly taken into consideration within selection assessments, requiring not only job activities to be assessed but also the ability to perform this task in a specific corporate culture (Ryan & Ployhart, 2014, Meyer et al., 2010). Therefore, person-environment fit is not only shown by person-job fit but also by person-organization fit (Segikuchi, 2004).

Interviews are also commonly used as a tool for selection (Chambers & Arnold, 2015). Interviews can take the form of unstructured, semi structured or structured interviews depending on the amount of job analysis conducted to guide the interviews. Evidence suggests that structured interviews have a higher predictive validity than assessments in assessment centers (Schmidt & Hunter, 1998). However, constructing structured interviews requires a thorough job and organization analysis, which in turn is cost extensive (König et al., 2010). In practice, interviews are often unstructured, having a significantly lower predictive validity than assessments in assessment centers (Schmidt & Hunter, 1998), and causing interpretation differences candidate results for the same application (Chambers & Arnold, 2015).

To address the issue of structuring interviews across participants for the same job, systems are developed which help to provide structure for interviews (Chambers & Arnold, 2015). These systems help to attach the answers to interview questions to a level of competency, as well as to standardize interview questions for candidates applying to the same job. Extensive interview training is conducted to help interviewers evaluate participants equally. However, even if such a system would work perfectly, several issues to validity remain when participant-interviewer interaction occurs. For example, the effect of social signals can influence interview outcomes (Bangerter et al., 2012). During an interview, participants are aware of being interviewed, and can therefore either try to show socially desirable behavior or non-consistent social signals which can lead to differences in interpretation of results (Jansen et al., 2012).

When using interviews, assessors are unable to observe actual workplace behavior. To address this issue, situational judgement tests (SJT) are often used. These tests require participants to judge behavior in a fictional work-related situation. The reaction to this fictional situation is used to form a judgement on actual workplace behavior (Ryan & Ployhart, 2014). However, the answers to these tests are often inaccurate predictions of actual workplace behavior, and individual differences in assessing situational demands influence outcomes of these questions (Jansen et al., 2012).

To address the dilemma between cost and test validity, HR has been developing various technologies for personnel selection (Ryan & Ployhart, 2014). For example, internet-based assessments have been used to be able to test a large number of selection participants simultaneously. The main advantage of using technology for selection is that using technology is considered to be more cost effective than traditional methods of assessment. In addition, computers are able to objectively adapt test style to participants, and measure results objectively (Tippins, 2011).

Most approaches in using technology for assessment have used a form of testing to determine knowledge, skills, abilities and other (Tippins, 2011), or to enhance predictive validity of current selection methods (Chambers & Arnold, 2015). Simulations have been used to specifically assess candidate behavior, although the use of simulation in candidate selection is still relatively rare (Ryan & Ployhart, 2014). Over the past few years, serious games are becoming an increasingly popular alternative for traditional selection methods (Fetzer, 2015).

Serious gaming

Using serious gaming with automatized observations is possibly a way to address the shortcomings of traditional selection methods. The term serious gaming is used to describe games used for other purposes than mere entertainment (Susi et al., 2007). In most cases, serious games are used to guide transformation processes or to facilitate cognitive learning processes (O'Neil et al., 2005). Over the past years, the popularity of using serious games for recruitment is increasing (Fetzer, 2015).

In the same way that serious games can help facilitate behavioral change, games can also be used to assess behavior in a structured way, thereby gaining deeper insight in competencies of participants (Schuller et al., 2013, Nacke et al., 2010). Automatically measuring competency-related behavior can eventually give an indication of the presence or absence of competency of participants.

Serious games can be a more effective alternative to traditional selection methods because 1. (digital) serious games can make use of computerized observations, making a more reliable way of collecting data (Tippins, 2011), 2. Serious games are able to simulate (elements of) a work environment, allowing participants to experience work situations before participating in a workplace environment (Wenzler, 2008), 3. Using computerized observation methods within serious games possibly allows more reliable assessments at little additional costs (Fetzer, 2015) and 4. Social desirability might be reduced because serious games allow a high level of engagement, making participants less aware of behavior during selection assessments (Shute & Kim, 2013).

Assessment in serious gaming is less obtrusive than traditional alternative to assessment (Westera et al., 2014, Mayer et al., 2013), because gaming absorbs participants in a state of play, thereby making participants less aware of behavior (Prensky, 2001). Although there are many definitions of play, in a broad sense play can be any activity that adds involvement and gives pleasure (Starbuck & Webster, 1991). Because play is a process which occurs naturally when playing games, and by nature utterly absorbs participants, a state of activity is created in which a person is less self-aware of its direct behavior (Prensky, 2001). Using serious gaming for competency assessment can provide a more unobtrusive way of assessment (Westera et al., 2014, Mayer et al., 2013).

Computers can be very consistent in measuring data, therefore making computer games a reliable way to gather data. Serious games are in most cases able to enter participants into a state called flow. A state of flow is created when a challenges given by a game are equal to participant skill level (Csikszentmihalyi, 2014). Flow causes continuous intrinsic motivation by challenging participants to achieve a higher skill level by providing continuous feedback (Prensky, 2001). This process causes participants to become less aware of the non-game environment, filtering out irrelevant thoughts and perceptions (Csikszentmihalyi, 2014). Participants unaware of being assessed are less likely to adjust behavior to achieve more favorable assessment outcomes (McCambridge et al., 2014).

In addition, serious games are able to simulate elements of workplace environments. This is facilitated by the ability of serous games to create memories of the future (Brandt, 2006, Susi et al., 2007, Wenzler, 2008). Creating memories of the future means that serious games are able to simulate or are a metaphor for real life scenarios, allowing scenarios to be experienced before these occur in a real life situation. These scenarios can be designed to facilitate participation in a real-time environment and promote awareness of behavior. For example, serious games have been used to train military in combat situations, or to train medical personnel in handling emergency situations (Susi et al., 2007).

Validating a serious game for selection assessment requires specific guidelines. For example, designers have to find a fit between structure and agency (Chin et al., 2009). Structure refers to the environment in which participants operate, whereas agency the choices that social actors make. Determining whether a particular choice is the product of an individual or from the environment is key when determining participant competence. In addition, validation of serious games requires not only the outcomes to be evaluated, but also the process of acquiring these outcomes (Belotti et al., 2013). Assessment based on only the outcomes is summative, whereas assessment based on the process of acquiring these outcomes is formative. Formative assessment in serious gaming is often referred to as stealth assessment (e.g. Shute, 2013, Mayer et al., 2013), because serious games are able to immerse the participant, revealing a more natural behavioral repertoire (Csikszentmihalyi, 2014).

Over the past decade, various serious games have been used for selection (Chin et al., 2009). Most of these serious games are designed for assessment of a specific purpose, such as the medical simulations, military training (Susi et al, 2007) or assessment in construction management (Mawdesley et al., 2011). Designing serious games around a specific purpose protects the validity of the assessment and allows game outcomes to be interpreted as participant skill level (Hummel et al., 2014, Gosen & Washbush, 2004).

Serious games have also been used to assess soft skills, such as professional skills (Laumer et al., 2012, Riedel & Hauge, 2011) or social competence (Hendrix et al., 2009). In these games, intended behavior is either measured by participant observations, or translated into in-game metrics to assess performance on these skills (Mayer et al., 2014, Crookall, 2010). These in-game metrics range from more simple metrics, such as time spent in game (Westera et al., 2014), and avoidable mistakes (Mayer et al., 2014). Avoidable mistakes are mistakes which have been made more than once such as dropping down the same cliff twice.

To some degree, previous studies have been successful at determining relevant behavior using only in game data. However, determining more complex behavior is still difficult due to challenges in measurement validity (Hummel et al., 2014, Chin et al., 2009) as well as a need for more accurate measurement methods (Belotti et al., 2013).

Automatically detecting behavior

Measures for directly detecting behavior can be found within social signal processing. This relatively new field aims to model, analyze and synthesize social behavior (Vinciarelli & Pentland, 2015). Social signals are acts or structures which influence the behavior of other individuals (Mehu & Scherer, 2012). Acts or structures can be either functional or informative. Functional components mainly include non-verbal acts, whereas informative components include more verbal aspects such as verbal expressions and emotions (Vinciarelli & Pentland, 2015). To be a social signal, acts and structures do not necessarily influence the behavior of another individual because this requires interpretation of another individual. Rather, acts or structures are labelled as social signals when these have the ability to convey information (Mehu & Scherer, 2012).

The measurement of social signals is inherent to selection assessment, as these constitute to an essential part of behavior. Various attempts have been made to measure behavior using various social signals. For example, Naim et al. (2015) use a combination of speaking style, word and facial analysis to predict job interview performance. Approaches designed more in the direction of competency-related behavior can be found within leadership. For example, Wang et al. (2012) developed a system which can detect leadership and cohesion in broadcast conversations. Similarly, Hadsell et al. (2012) use topic modelling to detect leadership in meetings. Common approaches to leadership modelling often include lexical features. Lexical features are speech-related measurements, such as pitch, speech speed and amount of speech and emotion analyses. Other approaches rely on turn-taking, where features like amount of speech segments, interruptions and speech duration are extracted (Vinciarelli et al., 2012). Most approaches are however limited in that these are uni-modal, using only one method of measurement (Naim et al., 2015, Zeng et al., 2007).

When looking at verbal communication, difficulties arise when measuring effective verbal communication from a social signal processing perspective. While the information content of a message can be the same, the interpretation of verbal messages can be dependent on culture, or even individual characteristics (Vinciarelli et al., 2012). In addition, the interpretation of a verbal message is dependent on the context in which the message is applied. Therefore, when looking at effective verbal communication, a significant difference exists between looking at verbal communication from a sender perspective or looking at verbal communication from a receiver perspective.

Listening is not always conveyed into social signals (Vinciarelli et al., 2012). Only when using either lexical utterances such as simply stating “yes” or “no”, or using back channels such as nodding, information content on listening is sent back to the speaker (McKneown et al., 2004). Similar to verbal communication, listening is context dependent, meaning that listening can only occur when another person is speaking (Vinciarelli et al., 2012).

While approaches at automatically detecting leadership have to successful, two restrictions apply in using these methods within a serious game for selection. One restriction is that specific methods have been trained on meeting or broadcast data (Vinciarelli et al., 2012). Using these on data during serious gaming might produce different results when for example participants are communicating through a computer screen. Second, there is no interaction between social signal processing and personnel selection literature. This means that measurement methods developed often do not connect to behaviors within competency frameworks. New approaches are required to achieve similar results within serious gaming.

To achieve the goal of automatic observations in serious gaming, more research is required on how to automatize behavior observations. Relatively few research exists on this topic. Therefore, exploratory research is required to determine what has to be done in order to automatize observation of competency-related behavior in the future. This study aims to find how competency-related behavior during a serious game can be automatically identified and measured to document and visualize the presence or absence of participant competency-related behavior.

Method

Research design

Literature offers a variety of automated measures for collecting data during game play which can be used to gather information on behavior. Some of these measures are found in game metrics include time taken (Mayer et al., 2014, Westera et al., 2014), avoidable mistakes and distance walked (Mayer et al., 2014). In addition, vocal speech and emotion recognition are used, as well as facial expression recognition. Although social signal processing signal processing offers a variety of other methods which might be useful when adapted to measure competency-related behavior, this research is limited to current commercialized methods because these methods are readily available. Game data is used to see to what extent game outcomes can be translated into game behavior. Previous approaches have mainly been successful with vocal speech recognition. However, including vocal and facial emotions in a multi-model approach might supplement prediction accuracy, achieving higher prediction levels.

TeamUp is used, a game in which 4 participants cooperate to complete five challenges. Participants are seated in the same room, each participant playing TeamUp from another laptop. During game play, participants have to cooperate in order to complete five challenges (Table 1). These challenges range from completing a maze to opening a door by standing on buttons. Participants have to

Challenge	Name	Description
1	Door puzzle	Participants need to navigate from their arrival dock to a closed door giving access to a cave. Entering the cave through the door requires coordinated action with two people needing to stand on two signs inside or outside to open the door and keep it open.
2	Tile puzzle	Participants have to find the correct path across a 8x8 tile maze. When a participant steps on a wrong tile, he will fall through and any of the team members can try again.
3	Maze puzzle	One participant stands high on platform where he has overview of three team members struggling to find the exit in a maze.
4	Bridge puzzle	The team needs to break up into various subgroups to solve small puzzles: a. entering a dark ruin where one team member leads with a flare and another needs to follow. One person needs to stay behind in the ruin standing on a sign. b. Two participants need to use their weight and distance to balance a bridge allowing them to climb onto a platform. One person needs to stay behind on the platform to stand on a sign. If and when four avatars stand on four signs dispersed throughout the level, a bridge to the next level is lowered.
5	Pillar puzzle	Team members alternate in leadership, trying to communicate and solve a series of four communication and coordinated action puzzles. Correctly solving one of the four puzzles opens a little bridge to the next puzzle where another team member becomes the leader of a similar, but more difficult, team challenge.

TABLE 1: OVERVIEW OF CHALLENGES IN TEAMUP (ADAPTED FROM MAYER ET AL., 2013)

collaborate to complete challenges. Each participant controls an avatar, which can be seen through the third person perspective. To avoid the influence of avatar choice on participant behavior (Lim & Reeves, 2009), characters are anonymized using hoodies. The game is controlled by the mouse and WASD keys to minimize skill advantages due to game competence (Mayer et al., 2013).

Behavior measurements and competencies are limited to behavior which can occur in a game. Four frequently used (sub)competencies are chosen which are present within the game used. These include interacting/ verbal communication, interacting/ listening, leading/taking initiative and leading/decision making. Translating competency to behaviors which occur in TeamUp requires competency to be broken down into traits demonstrating each competency, which again have to be broken down into examples of behaviors which demonstrate these traits specifically within TeamUp. These examples are competency related behaviors which demonstrate proficiency in the four competencies used. Competencies, traits and examples are shown in table 2.

To verify the frequency of behaviors during TeamUp, a four-point self-report and peer Likert scale survey is used. Using a self-report and a peer survey is more likely to give a complete picture of behaviors during play of TeamUp. Multiple opinions are required because engagement in serious games allows participants to become less aware of surroundings (Csikszentmihalyi, 2014) and the low amount of psychometric expertise among participants makes it difficult to accurately assess competency-related behavior. The survey contains five questions per competency. Each question is related to the presence and frequency of one behavioral example related to one of the four competencies. Participants self-report

behavior, as well as evaluate peer behavior. Results of the self-survey and the three peer surveys of a participants are averaged. Average behavior presence indicated by the averages of the survey is correlated with game data, vocal and facial recognition analysis.

Sample

Data was collected from a total of 72 participants. These participants are graduate students (N=49) or HR professionals (N=23). 54,1% of participants are male and 45,9% female. Average age of participants is 25,83 years. Because of constraints in subject availability, convenience sampling is used.

Instruments/measurements

This study uses in-game data measurements, vocal and facial recognition analysis. Game data includes total time taken and avoidable (repeated) mistakes. Speech analysis includes lexical analysis, word segment and relative volume identification, as well as emotion analysis. Transcription is conducted using Vocapia, an online tool for transcription (Vocapia, 2016). Automatic speech recognition is analyzed to find identification of speech length and speed. Speech segments are analyzed by a custom script, which detects segments at a minimum amplitude of 0,1 and a minimum of 3,125 seconds between segments. Vocal emotion analysis is conducted using Beyond Verbal, an application which is able to detect

Interacting/listening	**Behavioral traits (9)**	**Examples (5)**
Definition: Able to understand the essence from spoken words and stimulating others to try and get their message(s) across	- Listening actively - Doesn’t interrupt - Accurately hears what is said - Asking questions to clarify meaning - Understanding information via verbal expressions - Responds to reactions - Asking follow up questions - Establishes rapport - Tactfully choosing appropriate words	- Summarizing what has been discussed - Asking follow-up questions - Paraphrasing what has been discussed - Restating opinions - Letting peers finish sentences
Interacting/verbal communication	**Behavioral traits (6)**	**Examples (5)**
Definition: Able to express messages, ideas and opinions in a clear and transparent way which is easily understandable	- Speaks clearly - Talks at a calm pace and pays attention to reactions of peers - Speaks in plain language - Avoids jargon, uses simple language - Gets the idea of the message across - Able to provide clear instructions	- Testing whether message is understood - Clarifies issues with examples - Adapts communication style to audience - Engages others in discussion - Delivers messages with least words
Leading/decision making	**Behavioral traits (7)**	**Examples (5)**
Definition: Able to make timely and effective decisions	- Makes clear cut decisions - Recognizes the importance of having necessary information to make sound decisions - Acts quickly and decisively - Assesses options during decision making process - Recognizes trade-offs - Chooses the appropriate action - Involves others	- Making decisions based on factual information - Making decisions based on experience - Making decisions based on judgments - Consults before coming to a decision - Generates alternatives
Leading/taking initiative	**Behavioral traits (4)**	**Examples (5)**
Definition: Able to spot chances and act properly, having a proactive attitude	- Undertakes unrequested action; seizes chances and opportunities - Actively seeks needed information - Has a pro-active attitude - Takes the lead - Keeps the initiative despite obstacles	- Takes initiative - Is involved in discussion - Leads discussion - Comes up with examples - Seeks needed information to solve issues

TABLE 2: COMPETENCIES, TRAITS AND EXAMPLES USED

valence, arousal and temper in individual segments of speech (Beyond Verbal, 2014). Video analysis is conducted using FaceReader, an application which uses facial recognition to detect and analyze appearance of six universally accepted emotions: angry, sad, happy, disgust, surprise, anxiety as well as neutral (Ekman, 1970). In addition, FaceReader is able to measure valence and arousal (Lewinsky et al., 2014). A full overview of automatic variables measured can be found in table 3.

A four point Likert survey is used (Appendix 2) to verify the frequency of behavior during game play. Participants self-report behavior, as well as the behavior of three peers. Scores of the three peer surveys and the self-report survey are averaged to determine the presence and frequency of competency-related behavior. Using a four point scale prevents socially desirable answers (Bertram, 2007, Garland, 1991) and retention due to survey length. To allow scaling, the survey is adapted to include scale variables (Hamby & Levine, 2016). In addition, the frequency of behaviors is included to prevent subjective judgement (Bertram, 2007). Answers range from “1. never (0)” to “4. always (>10)”. Participants can indicate “0. not applicable” if no reliable judgement can be given.

Procedure

Participants are randomly assigned in teams of four. Before playing the game, participants are asked to complete the game as efficiently and effectively as possible. During game play, audio and video are recorded using a headset and laptop webcam. Participants are aware that video and audio is being recorded, although being unaware of the purpose of recording. This prevents participants from alternating behavior, and therefore protects content validity (Bryman & Bell, 2015). Directly after playing the game, participants complete the self-report and peer survey. Audio and video recorded is synchronized to provide similar measurement timings for each of the four participants recorded in a team. Audio and video is decomposed into one segment for each challenge, yielding a total of five segments per participant recorded. All audio and video segments are rendered at

	Variable	Definition	Scale
Game data	**Time taken**	Time spent to complete game (minutes)	Continuous
	Avoidable mistakes	Mistakes made more than once	Continuous
	Distance	Distance covered in game (meters)	Continuous
Audio analysis	**Valence (speech)**	Positivity of speech	0-100
	Arousal (Speech)	Alertness of participant	0-100
	Temper	Transitory emotional state (temperament)	0-100
	Volume	Volume relative to participant	participant average=1
	% talking	Percentage of total time speaking	0-1
	Words per minute	Amount of words per minute of game time	Continuous
	Sentences per minute	Amount of sentences per minute of game time	Continuous
	Word length	Average word duration (seconds)	Continuous
	Interruptions	Interruptions per minute	Continuous
	Agree/minute	Agreements expressed per minute. Agreement is either stating "yes" or "ok".	Continuous
	Disagree/minute	Disagreements expressed per minute. Agreement is stating "no".	Continuous
Facial expression analysis	**Neutral**	No significant emotion is shown	0-1
	Happy		0-1
	Sad		0-1
	Anxiety		0-1
	Surprise		0-1
	Scared		0-1
	Disgusted		0-1
	Valence (video)	Positivity of facial expression	0-1
	Arousal (video)	Degree of focus shown in face	0-1

TABLE 3: OVERVIEW OF AUTOMATIC MEASUREMENTS

similar quality, 30 frames per second and 512 MB per second for video, 16 kb per second for audio.

Results

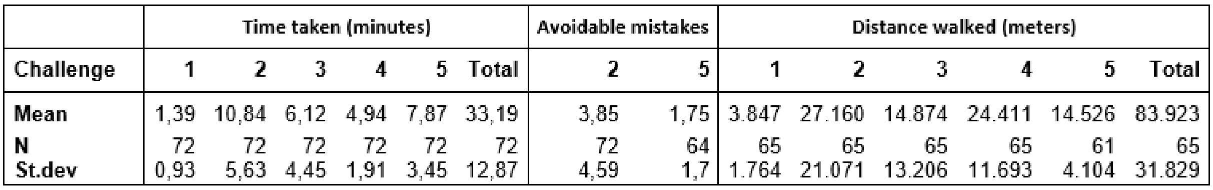
Descriptive results

Descriptive results for the survey can be found as an appendix. Descriptive results for game data, audio and facial recognition analysis are shown in table 4, 5 and 6 respectively. These results are used to determine how participants play TeamUp and give context to behavior prediction models presented later in this section

Average time taken for playing TeamUp is 33,19 minutes. The standard deviation of 12,87 minutes indicates a considerable spread across the participants For example, time taken for completing TeamUp ranges from 12,9 minutes to 56,43 minutes. There is a high correlation between age and game time (p: 0,683), indicating that participants with higher ages take longer to complete TeamUp. Male participants complete the game faster than female participants (p=0,786). Both correlations are significant at the 0,01 level.

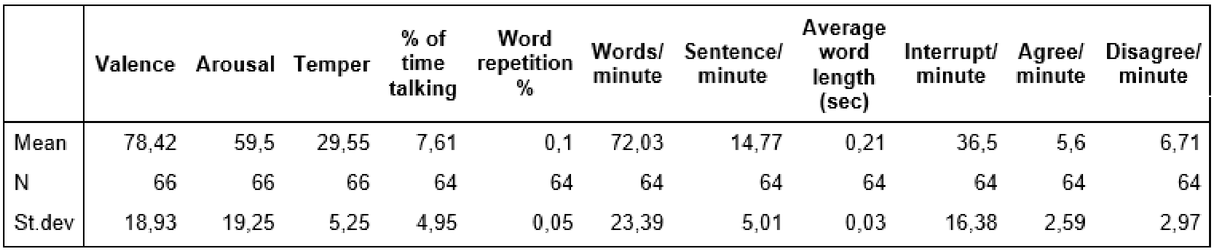
High valence and arousal values indicate that participants experience playing TeamUp as positive, and are engaged during game play. Words and sentences are often short. Participants use an average of 4,88 words per sentence, from which a large number of sentences consist of either expressing confirmation (5,6/minute) or disagreement (6,71 per minute). A large number of words is spoken during communication of other participants, indicated by an average of 36,50 words interrupted per minute.

The main emotions shown during the game are neutral and happiness, with a mean of 0,586 and 0,151 respectively. Emotions among participants vary considerably, indicated by the large standard deviations among all emotions. Valence levels measured by face



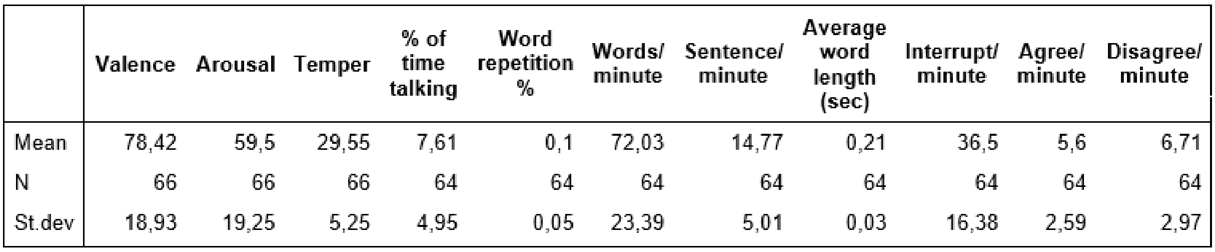
	Time taken (minutes)						Avoidable mistakes		Distance walked (meters)					
Challenge	1	2	3	4	5	Total	2	5	1	2	3	4	5	Total
Mean	1,39	10,84	6,12	4,94	7,87	33,19	3,85	1,75	3.847	27.160	14.874	24.411	14.526	83.923
N	72	72	72	72	72	72	72	64	65	65	65	65	61	65
St.dev	0,93	5,63	4,45	1,91	3,45	12,87	4,59	1,7	1.764	21.071	13.206	11.693	4.104	31.829

TABLE 4: DESCRIPTIVE RESULTS FOR GAME DATA



	Valence	Arousal	Temper	% of time talking	Word repetition %	Words/ minute	Sentence/ minute	Average word length (sec)	Interrupt/ minute	Agree/ minute	Disagree/ minute
Mean	78,42	59,5	29,55	7,61	0,1	72,03	14,77	0,21	36,5	5,6	6,71
N	66	66	66	64	64	64	64	64	64	64	64
St.dev	18,93	19,25	5,25	4,95	0,05	23,39	5,01	0,03	16,38	2,59	2,97

TABLE 5: DESCRIPTIVE RESULTS FOR VOCAL ANALYSIS



	Valence	Arousal	Temper	% of time talking	Word repetition %	Words/ minute	Sentence/ minute	Average word length (sec)	Interrupt/ minute	Agree/ minute	Disagree/ minute
Mean	78,42	59,5	29,55	7,61	0,1	72,03	14,77	0,21	36,5	5,6	6,71
N	66	66	66	64	64	64	64	64	64	64	64
St.dev	18,93	19,25	5,25	4,95	0,05	23,39	5,01	0,03	16,38	2,59	2,97

TABLE 6: DESCRIPTIVE RESULTS FOR FACIAL ANALYSIS

recognition are also high (mean: 0,612), whereas arousal values measured by facial recognition are lower than arousal measured by vocal analysis. Correlating valence and arousal values from both speech and facial recognition reveals that valence values between both measures show no significant correlation (s: 0,34, p: 0,12), as opposed to arousal (p: 0,32, s: 0,01).

Modelling

Survey results are analyzed find extreme cases (formula $\sqrt{(multiple\ mode - \boldsymbol{x})^{\boldsymbol{2}}} < 1{,}5$). Extreme cases are removed to improve reliability. Remaining survey results are analyzed using Cronbach's α to find the level of agreement between different answers on the same behavior for the same participant. Averages are calculated of remaining survey results to determine the frequency of competency-related behavior for each participant.

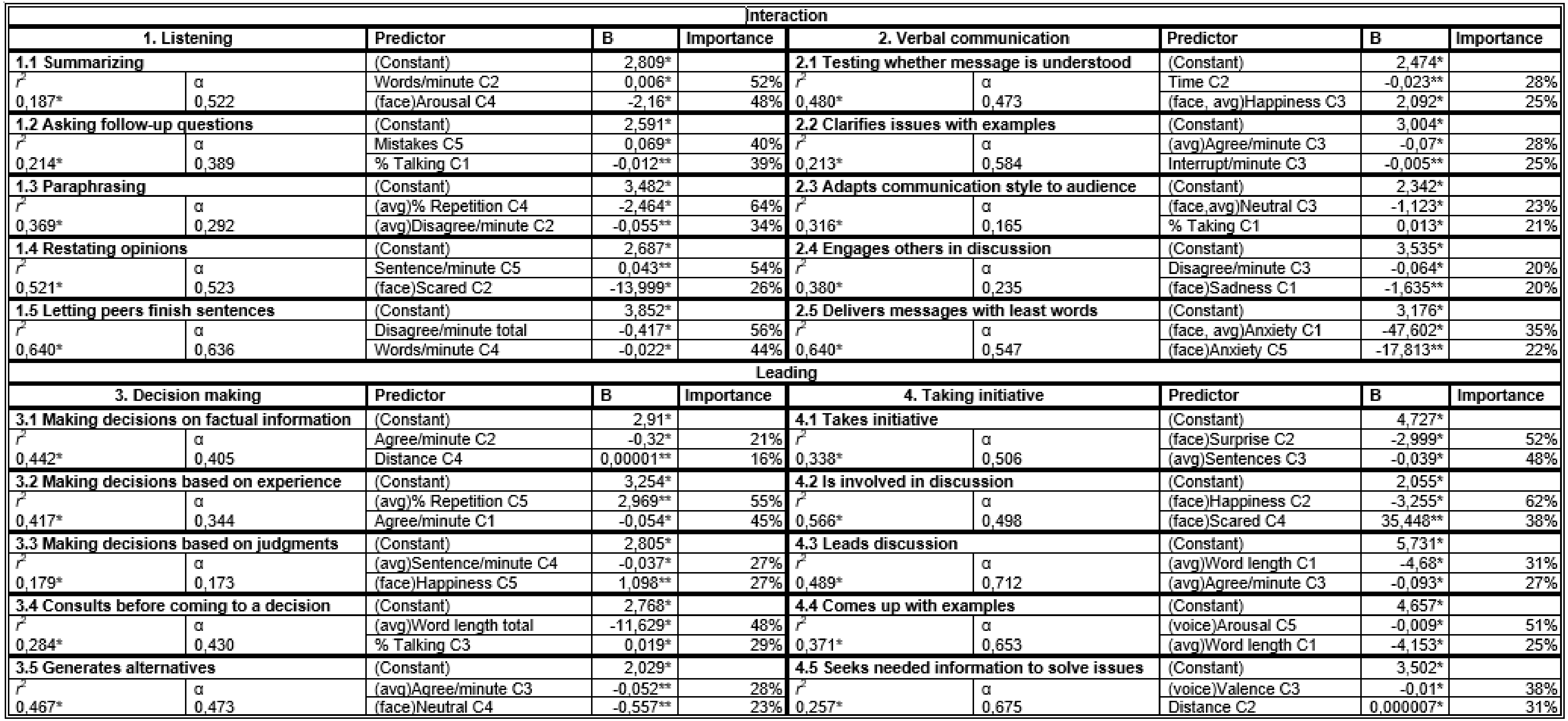
Results of automated measurements (game data, vocal and facial expression analysis) are correlated with averages of survey results. Correlating results of automated measurement methods with averages of self-report and peer survey data to be used for multiple regression reveals a complex network of variables. To simplify the network for each of the competency-related behaviors, only significant correlations are added to each behavior model. Insignificant slope determinants are removed individually, until a significant model exists which contains only significant slope variables. This process results in multiple regression models predicting each competency-related behavior using 2 to 11 automated measurements. Results are shown in table 7.

Competency-related behavior for verbal communication and taking initiative can be best predicted using game data, vocal and facial recognition analysis within TeamUp, with an average r^2 of 0,406. Third is listening, with an average r^2 of 0,386. Fourth is taking initiative with an average r^2 of 0,358.. Models with higher explanatory powers are most likely also the models using most predictors (p: 0,687, s: 0,001).

Average α among all survey results is considerably low (average α: 0,472). Highest consensus is found among behaviors relating to taking initiative (average α: 0,608) indicating these behaviors are less difficult to assess with a self-report and peer survey. Lowest consensus is found among decision making (average α: 0,365).

Highest prediction and reliability levels within behaviors related to listening are found in restating opinions (r^2: 0,521, α: 0,523) and letting peers finish sentences (r^2: 0,640, α: 0,636). The largest predictor of restating opinions is sentence/minute in challenge 5 (54%), which is a considerable difference to the second best predictor scaredness in challenge 2 (26%). Although the amount of sentences in challenge 5 can contribute towards restating opinions, it is difficult to see why this specifically applies to challenge 5. Similarly, while a low amount of disagreements in total and words in challenge 4 can have a connection to letting peers finish sentences, there are difficulties in finding a direct relationship between this competency-related behavior and predictor measurements.

For verbal communication, highest prediction and reliability levels within behaviors are found in



Interaction									
1. Listening		Predictor	B	Importance	2. Verbal communication		Predictor	B	Importance
1.1 Summarizing		(Constant)	2,809*		2.1 Testing whether message is understood		(Constant)	2,474*	
r^2	α	Words/minute C2	0,006*	52%	r^2	α	Time C2	-0,023**	28%
0,187*	0,522	(face)Arousal C4	-2,16*	48%	0,480*	0,473	(face, avg)Happiness C3	2,092*	25%
1.2 Asking follow-up questions		(Constant)	2,591*		2.2 Clarifies issues with examples		(Constant)	3,004*	
r^2	α	Mistakes C5	0,069*	40%	r^2	α	(avg)Agree/minute C3	-0,07*	28%
0,214*	0,389	% Talking C1	-0,012**	39%	0,213*	0,584	Interrupt/minute C3	-0,005**	25%
1.3 Paraphrasing		(Constant)	3,482*		2.3 Adapts communication style to audience		(Constant)	2,342*	
r^2	α	(avg)% Repetition C4	-2,464*	64%	r^2	α	(face,avg)Neutral C3	-1,123*	23%
0,369*	0,292	(avg)Disagree/minute C2	-0,055**	34%	0,316*	0,165	% Taking C1	0,013*	21%
1.4 Restating opinions		(Constant)	2,687*		2.4 Engages others in discussion		(Constant)	3,535*	
r^2	α	Sentence/minute C5	0,043**	54%	r^2	α	Disagree/minute C3	-0,064*	20%
0,521*	0,523	(face)Scared C2	-13,999*	26%	0,380*	0,235	(face)Sadness C1	-1,635**	20%
1.5 Letting peers finish sentences		(Constant)	3,852*		2.5 Delivers messages with least words		(Constant)	3,176*	
r^2	α	Disagree/minute total	-0,417*	56%	r^2	α	(face, avg)Anxiety C1	-47,602*	35%
0,640*	0,636	Words/minute C4	-0,022*	44%	0,640*	0,547	(face)Anxiety C5	-17,813**	22%
Leading									
3. Decision making		Predictor	B	Importance	4. Taking initiative		Predictor	B	Importance
3.1 Making decisions on factual information		(Constant)	2,91*		4.1 Takes initiative		(Constant)	4,727*	
r^2	α	Agree/minute C2	-0,32*	21%	r^2	α	(face)Surprise C2	-2,999*	52%
0,442*	0,405	Distance C4	0,00001**	16%	0,338*	0,506	(avg)Sentences C3	-0,039*	48%
3.2 Making decisions based on experience		(Constant)	3,254*		4.2 Is involved in discussion		(Constant)	2,055*	
r^2	α	(avg)% Repetition C5	2,969**	55%	r^2	α	(face)Happiness C2	-3,255*	62%
0,417*	0,344	Agree/minute C1	-0,054*	45%	0,566*	0,498	(face)Scared C4	35,448**	38%
3.3 Making decisions based on judgments		(Constant)	2,805*		4.3 Leads discussion		(Constant)	5,731*	
r^2	α	(avg)Sentence/minute C4	-0,037*	27%	r^2	α	(avg)Word length C1	-4,68*	31%
0,179*	0,173	(face)Happiness C5	1,098**	27%	0,489*	0,712	(avg)Agree/minute C3	-0,093*	27%
3.4 Consults before coming to a decision		(Constant)	2,768*		4.4 Comes up with examples		(Constant)	4,657*	
r^2	α	(avg)Word length total	-11,629*	48%	r^2	α	(voice)Arousal C5	-0,009*	51%
0,284*	0,430	% Talking C3	0,019*	29%	0,371*	0,653	(avg)Word length C1	-4,153*	25%
3.5 Generates alternatives		(Constant)	2,029*		4.5 Seeks needed information to solve issues		(Constant)	3,502*	
r^2	α	(avg)Agree/minute C3	-0,052**	28%	r^2	α	(voice)Valence C3	-0,01*	38%
0,467*	0,473	(face)Neutral C4	-0,557**	23%	0,257*	0,675	Distance C2	0,000007*	31%

Table 7: Prediction levels of competency-related behavior and most important model variables predicting specific behaviors. Models and slopes marked with * and ** are significant at the 0,01 and 0,05 levels respectively. Variables marked with (avg) are measured as distance from average ($\sqrt{(\bar{x}-x)^2}$), variables marked with (face) or (voice are emotions measured trough facial and vocal analysis respectively. Variables marked with C (e.g. C1) are specifically measured during one of the five challenges in TeamUp, whereas variables marked with "total" are measured during the entire length of playing TeamUp. Only the two most significant predictors are shown for each model. Predicted variable scores range from 1 to 4.

testing whether the message is understood (r^2: 0,480, α: 0,473) and delivering messages with least words (r^2: 0,640, α: 0,547). The only prediction of competency-related behavior which to some extent can be explained logically, is a lower time taken for challenge 2 predicting understanding of delivered messages, as challenge 2 requires participants to communicate the correct path. Testing if a message is understood can contribute towards a lower time for this challenge.

Within the competency of decision making, high prediction and reliability levels within behaviors are found in making decisions on factual information (r^2: 0,442, α: 0,405) and generating alternatives (r^2: 0,467, α: 0,473). Although agreeing less often in challenge 2 may indicate participants listening less to other participants, thereby using more factual information for progression, there is no indication as to why this would not be more important in other challenges as well.

Lastly, behaviors related to the competency of taking initiative, high prediction and reliability levels are found in being involved in (r^2: 0,566, α: 0,489) and leading discussions (r^2: 0,489, α: 0,712). In both behaviors facial emotions play a significant role in predicting being involved in discussion, while leading a discussion is mainly predicted by word length and the number of agreements. These findings could indicate that participants leading discussions convey less different emotions than participants participating in discussions.

Reliability of results is considerably low, making accurate predictions difficult to generalize towards new sample groups as well as other situations. Low reliability might indicate either than TeamUp immerses participants into the game, making accurate behavior predictions difficult or than some expertise of behaviors is required to make an accurate prediction on competency-related behaviors. Looking at predictions and predictors of competency-related behavior, very few relationships between automated observations and survey behavior predictions can be logically explained. Although some relationships are logically explained, direct relationships between automated measurements and survey behavior predictions are hard to establish.

In addition, while some logical relationships between survey data and automated measurements are confirmed by correlation analysis, these did not show up as most important predictors in the regression analysis. For example, lettings peers finish sentences increased as the number of interruptions decreased (p: -0,358, s<0,01) but this variable was not significant when added to a multiple regression model with other significant predictors. Higher correlations are expected if measured interruptions are to detect how often participants let peers finish sentences.

Discussion

This section evaluates results to determine what should be done in the future to be able to find how competency-related behavior during a serious game can be automatically identified and measured to document and visualize the presence or absence of participant competency-related behavior. Evaluating results leads to several propositions, which can be used to guide future research.

Although mainly verbal communication and decision making can to a large extent be predicted within TeamUp using game data, vocal and facial recognition analysis, a lack of logical explanation for predicted variables prevent any findings to be generalized towards other situations. The main reason for the lack of generalizability is the difficulty in directly observing behavior. Although interruptions can be measured (semi)directly using commercialized automated software testing more complex behaviors requires more advanced software. For example, asking follow-up questions is difficult to measure because within software used for this research few measures exists which can directly identify questions based on vocal analysis. The number of participants also limits the extent to which statistics can be used. A larger sample size might have eliminated the chance that non-significant variables are seen as significant predictors of behavior and increased the chance that a significant predictor of behavior is included as the most important predictor of behavior. In addition, a larger sample size might have provided the opportunity for more complex prediction models to be developed.

P1: Research on automatic detection of competency-related behavior should use larger sample sizes to increase internal validity.

Using self and peer surveys during serious games is difficult because of the engagement which games provide. Average Cronbachs α of survey results is 0,472 after removing extreme values, indicating that either competency-related behaviors are too complex to assess after playing a serious game, or that the engagement provided by TeamUp allows participants to be less aware of other elements of the environment. In addition, difficulties arise when trying to measure the exact numbers of behavior occurrence. Using a four-point scale only gives a general indication of behavior frequency, and does not allow automated methods of behavior measurement to be verified directly. Allowing participants to look into their own game play after participating in play of a serious game or using observations by assessors with psychometric expertise can provide a more accurate prediction of the exact number behavior has occurred.

P2: Manual observations are likely to be a more accurate method of validating automated behavior measurements within serious games than peer and self-surveys.

Future attempts in measuring behavior should focus on translating game outcomes to a conclusion on behavior. This possibility was not present in TeamUp due to the high correlations between gender and game time, as well as age and game time. In order to be able to translate game behavior to conclusions on job behavior, research should be conducted to assess similarities between game behavior and job behavior. Current research on this subject exists, but focuses mainly on similarities between behavior in entertainment games and real world behavior. Future research can address this issue by focusing on the similarities between participant behavior in a selection game and job environment behavior. Serious games often hold the assumption that reality is similar to the situation within the serious game. Using serious games for selection requires these assumptions to be validated in order to be able to predict person-environment fit.

P3: More extensive validation of serious games is required when using serious games for selection purposes to make sure that more competent persons achieve higher outcomes.

The approach of this research has taken less consideration to context of behaviors. In line with social signal processing and serious gaming literature, confirming evidence is found that behavior is context dependent. If an action were to be measured directly, this action would convey a different meaning within a different situation. For example, summarizing might be useful in a situation where another participant tries to convey useful information, but might be less useful when other participants are conveying less useful information. Direct measurement of behavior requires these nuances to be taken into consideration to develop more accurate measurement methods, either by predetermining a context in a specific moment in the game, or by finding ways of measuring context.

P4: Behavior in serious games is context dependent. Detecting behavior requires either context to be built in during a specific moment in the game, or a method of combining context and behavior.

Although serious gaming, selection and assessment literature are becoming increasingly connected, few effort exists which connects this literature to the measurement literature in social signal processing. Future research should focus on creating connections between demand for measurements and development of automated measurements by attempting to automatize social signals related to behavior within competencies.

P5: Personnel selection and serious gaming should establish to social signal processing literature to be able to automatize observations of relevant competency-related behavior.

Conclusion

This research has explored how competency-related behavior during a serious game can be automatically identified and measured to document and visualize the presence or absence of participant competency-related behavior. Results indicate that although using a variety of modern measurement methods allow to a large extent prediction of several competency-related behaviors, the extent to which these predictions of behaviors can be logically explained by automated measurements predicting these behaviors is limited. Using game data to determine behavior is difficult for TeamUp, as showing a high degree of competency-related behavior does not result in more desirable game outcomes. More accurate methods for measuring competency-related behavior are likely to be found by designing serious games to validate outcomes to proficiency levels in competency. In addition, using a specific combination of measurement techniques and more complex models is more likely to yield to direct measurements on the frequency of competency-related behavior. Connecting social signal processing literature to current collaborations between personnel selection and serious gaming literature is likely to lead to more accurate and relevant measurement methods for measuring competency-related behaviors.

Acknowledgements

Beyond Verbal, Noldus and Vocapia are acknowledged for providing their software for usage during this project. No conflict of interest has occurred in the usage of this software.

References

Bangerter, A., Roulin, N., & König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology*, *97*(4), 719.

Bartram, D. (2005). The Great Eight competencies: a criterion-centric approach to validation. *Journal of applied psychology*, *90*(6), 1185.

Bartram, D. (2012). The SHL universal competency framework. *Surrey, UK: SHL White Paper*.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction*, *2013*, 1.

Bertram, D. (2007). Likert scales.

Beyond Verbal (2014) Deciphering the Intonation Code for Emotions Analytics

Brandt, E. (2006). Designing exploratory design games: a framework for participation in participatory design?. In *Proceedings of the ninth conference on Participatory design: Expanding boundaries in design-Volume 1* (pp. 57-66). ACM.

Bryman, A., & Bell, E. (2015). *Business research methods*. Oxford university press.

Cheney, P. H., Hale, D. P., & Kasper, G. M. (1990). Knowledge, skills and abilities of information systems professionals: past, present, and future. *Information & Management*, *19*(4), 237-247.

Chin, J., Dukes, R., & Gamson, W. (2009). Assessment in Simulation and Gaming A Review of the Last 40 Years. *Simulation & Gaming*, *40*(4), 553-568.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83-117.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. *Simulation & gaming*, *41*(6), 898-920.

Ekman, P., & Keltner, D. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, *8*(4), 151-158.

Fetzer, M. (2015). Serious games for talent selection and development. *TIP: The Industrial-Organizational Psychologist*, *52*, 117-125.

Galley, M., McKeown, K., Hirschberg, J., & Shriberg, E. (2004, July). Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* 669-677. Association for Computational Linguistics.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing bulletin*, *2*(1), 66-70.

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in Psychological Assessment: Recognizing the people behind the data. *Psychological Assessment*, *23*(3), 656.

Hamby, T., & Levine, D. S. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, 0146621615597961.

Hauge, J. B. (2011). State of the art of serious games for business and industry. In *Concurrent Enterprising (ICE), 2011 17th International Conference on concurrent enterprising* 1-8. IEEE.

Hendrix, K., van Herk, R., Verhaegh, J., & Markopoulos, P. (2009, June). Increasing children's social competence through games, an exploratory study. In *Proceedings of the 8th International Conference on Interaction Design and Children* 182-185. ACM.

Hummel, H., Nadolski, R., Joosten-ten Brinke, D., & Baartman, L. (2014). Validation of game scenarios for the assessment of professional competence: Development of a serious game for system managers in training.

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology*, *98*(2), 326.

Kaslow, N. J., Rubin, N. J., Bebeau, M. J., Leigh, I. W., Lichtenberg, J. W., Nelson, P. D., ... & Smith, I. L. (2007). Guiding principles and recommendations for the assessment of competence. *Professional Psychology: Research and Practice*, *38*(5), 441.

König, C. J., Klehe, U. C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, *18*(1), 17-27.

Laumer, S., Eckhardt, A., & Weitzel, T. (2012). Online Gaming to Find a New Job—Examining Job Seekers' Intention to Use Serious Games as a Self-Assessment Tool. *Zeitschrift für Personalforschung/German Journal of Research in Human Resource Management*, 218-240.

Le Deist, F. D., & Winterton, J. (2005). What is competence?. *Human resource development international*, *8*(1), 27-46.

Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, *7*(4), 227.

Lim, S., & Reeves, B. (2009). Being in the game: Effects of avatar choice and point of view on psychophysiological responses during play. *Media Psychology*, *12*(4), 348-370.

Mawdesley, M., Long, G., Al-Jibouri, S., & Scott, D. (2011). The enhancement of simulation based learning exercises through formalised reflection, focus groups and group presentation. *Computers & Education*, *56*(1), 44-52.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., ... & Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology*, *45*(3), 502-527.

Mayer, I., van Dierendonck, D., van Ruijven, T., & Wenzler, I. (2013). Stealth assessment of teams in a digital game environment. In *Games and Learning Alliance* (pp. 224-235). Springer International Publishing.

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology*, *67*(3), 267-277.

Mehu, M., & Scherer, K. R. (2012). A psycho-ethological approach to social signal processing. *Cognitive processing*, *13*(2), 397-414.

Meyer, J. P., Hecht, T. D., Gill, H., & Toplonytsky, L. (2010). Person–organization (culture) fit and employee commitment under conditions of organizational change: A longitudinal study. *Journal of Vocational Behavior*, *76*(3), 458-473.

Nacke, L. E., Drachen, A., & Göbel, S. (2010). Methods for evaluating gameplay experience in a serious gaming context. *International Journal of Computer Science in Sport*, *9*(2), 1-12.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015, May). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on Acoustics, Speech and Signal Processing* 1, 1-6. IEEE.

Nakamura, J., & Csikszentmihalyi, M. (2014). The concept of flow. In *Flow and the Foundations of Positive Psychology* 239-263. Springer Netherlands.

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Cirriculum Journal*, *16*(4), 455-474.

Prensky, M. (2001). Fun, play and games: What makes games engaging. *Digital game-based learning*, *5*, 1-05.

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*(1), 114.

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual review of psychology*, *65*, 693-717.

Ryan, A. N. N., McFarland, L., & SHL, H. B. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology*, *52*(2), 359-392.

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

Schuller, B. W., Dunwell, I., Weninger, F., & Paletta, L. (2013). Serious gaming for behavior change: The state of play. *IEEE pervasive computing*, (3), 48-55.

Sekiguchi, T. (2004). Person-organization fit and person-job fit in employee selection: A review of the literature. *Osaka keidai ronshu*, *54*(6), 179-196.

Shute, V. J. (2009). Simply assessment.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In *Handbook of research on educational communications and technology* 311-321. Springer New York.

Starbuck, W. H., & Webster, J. (1991). When is play productive?. *Accounting, Management and Information Technologies*, *1*(1), 71-90.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games: An overview.

Tippins, N. T. (2011). Overview of Technology-Enhanced Assessments. *Technology-enhanced assessment of talent*, 1-18.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing, IEEE Transactions on*, *3*(1), 69-87.

Vinciarelli., A. & Pentland., A. (2015) New Social Signals in a New Interaction World: The Next Frontier for Social Signal Processing, *IEEE Systems, Man and Cybernetics Magazine*, *1*(2)10-17

Vocapia (2016). Vocapia speech to text software, *http://www.vocapia.com/*

Wang, W., Precoda, K., Hadsell, R., Kira, Z., Richey, C., & Jiva, G. (2012, March). Detecting leadership and cohesion in spoken interactions. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on Acoustics, Speech and Signal Processing* 5105-5108 IEEE.

Wenzler, I. (2008). Is your simulation game blue or green. *Caluwé, L de, G. Hofstede & V. Peters. Why do games work. In search for the active substance. Deventer: Kluwer*.

Westera, W., Nadolski, R., & Hummel, H. (2014). Serious Gaming Analytics: What Students Log Files Tell Us about Gaming and Learning.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(1), 39-58.

Appendix: survey overview

Number	**Question***	**Mean**	**St. dev.**	**Alpha**
Q1.1	... someone who summarizes what has been discussed?	2,54	0,40	0,522
Q1.2	... someone who asks follow-up questions?	2,59	0,35	0,389
Q1.3	... someone who paraphrases what has been discussed?	2,59	0,35	0,292
Q1.4	... someone who restates the opinion of others?	2,45	0,42	0,523
Q1.5	... someone who lets other people finish their sentences?	3,00	0,57	0,636
Q2.1	... someone who tests whether the message is properly understood?	2,69	0,40	0,473
Q2.2	... someone who clarifies issues/situations using the right examples?	2,62	0,38	0,584
Q2.3	... someone who adapts his communication style, depending on the audience and situation?	2,72	0,37	0,165
Q2.4	... someone who engages others in a discussion?	2,77	0,49	0,235
Q2.5	... someone who delivers messages using the least words as possible?	2,61	0,41	0,547
Q3.1	... someone who makes decisions based on the analysis of factual information?	2,78	0,41	0,405
Q3.2	... someone who makes decisions based on the analysis of experience?	2,84	0,41	0,344
Q3.3	... someone who makes decisions based on the analysis of judgments?	2,75	0,36	0,173
Q3.4	... someone who consults with others before coming to a decision?	2,67	0,39	0,430
Q3.5	... someone who generates alternatives?	2,78	0,40	0,473
Q4.1	... someone who takes the initiative?	2,72	0,53	0,506
Q4.2	... someone who is involved in discussions?	2,84	0,40	0,498
Q4.3	... someone who leads discussions?	2,66	0,55	0,712
Q4.4	... someone who comes up with examples to solve (un)expected issues?	2,63	0,48	0,653
Q4.5	... someone who actively seeks the needed information to solve (un)expected issues?	2,72	0,43	0,675

*… is replaced by "Are you" in the self-report survey and "Is your peer" in the peer survey. Survey scale ranges from 1 (never) to 4 (always).