

# Master Thesis

EXPLORING THE ADDED VALUE OF OBSERVATIONAL METHODS IN SURVEY-BASED  
TEAM PSYCHOLOGICAL SAFETY RESEARCH

Waria Gankema | s1758438

Supervisors: Dr. D.H. Van Dun and Prof. dr. C.P.M. Wilderom

November, 2020

**UNIVERSITY  
OF TWENTE.**

## ABSTRACT

Team psychological safety describes the believe of all team members within a team that it is safe to take interpersonal risks, such as voicing concerns, admitting to mistakes and raising ideas, without having to expect negative repercussions for the behaviour. The construct is typically measured using self-report survey measures which have inherent limitations, such as self-report and non-response bias. It is proposed that using observational research methods in concert with self-report measures can counteract these limitations. To this end, this research explores the added value of an observation scheme for measuring team psychological safety.

An existing psychological safety observation scheme for observation of team meetings is refined and then used in four studies with different team samples ( $n=4$ ,  $n=7$ ,  $n=6$ ,  $n=1$ ). The data from the observation scheme is combined with survey and qualitative data. Observations are conducted, both naked-eye and with the help of The Observer XT software, and differences in these approaches are discussed.

It has been found that observational research methods can support the findings of surveys through both triangulation and crystallization. The observational results show that there are distinct meeting behaviours that are related to psychological safety. For example, the behaviours *Agreeing*, *Asking for ideas, help or solutions*, *Sharing future plans* and *Providing information* occur significantly more often in teams with higher psychological safety. It has been found that computer-aided observations enrich data analysis but cost substantially more time to analyse than naked-eye observations. Limitations of the research are discussed and avenues for future research are proposed.

*Keywords:* Psychological safety, Observations, Mixed-methods research

## TABLE OF CONTENTS

<b>Abstract .....</b>	<b>1</b>
<b>Introduction.....</b>	<b>3</b>
<b>Theoretical framework .....</b>	<b>5</b>
<i>Psychological Safety.....</i>	<i>5</i>
<i>Observational Research Methods.....</i>	<i>10</i>
<b>Methodology .....</b>	<b>15</b>
<i>Overall Research Design .....</i>	<i>15</i>
<i>Pilot study .....</i>	<i>17</i>
<i>Study 1 .....</i>	<i>18</i>
<i>Study 2 .....</i>	<i>20</i>
<i>Study 3 .....</i>	<i>23</i>
<i>Study 4 .....</i>	<i>25</i>
<b>Findings .....</b>	<b>26</b>
<i>Pilot study .....</i>	<i>26</i>
<i>Study 1 .....</i>	<i>26</i>
<i>Study 2 .....</i>	<i>31</i>
<i>Study 3 .....</i>	<i>42</i>
<i>Study 4 .....</i>	<i>51</i>
<i>Cross-study comparison .....</i>	<i>55</i>
<b>Discussion .....</b>	<b>57</b>
<i>Theoretical implications.....</i>	<i>58</i>
<i>Practical implications.....</i>	<i>61</i>
<i>Limitations and future research.....</i>	<i>61</i>
<b>Conclusion .....</b>	<b>63</b>
<b>References.....</b>	<b>65</b>
<b>Appendix .....</b>	<b>70</b>

## INTRODUCTION

In today's society, teams have become a prevalent form of organizing work (Delgado Pina, Romero Martinez, & Gomez Martinez, 2008; Kostopoulos & Bozionelos, 2011; Salas, Cooke, & Rosen, 2008). This has been owed to the increasing complexity and difficulty of tasks in organizations, to the extent that they are not executable by single individuals anymore (Salas et al., 2008). Moreover, using teams can increase flexibility and adaptation of organizations (Delgado Pina et al., 2008). A team is defined as a group of individuals who work, collectively and interdependently, on organizationally relevant tasks to achieve a common goal (Kozlowski & Ilgen, 2006). To achieve that common goal, it is important that teams perform well.

Mathieu, Hollenbeck, Van Knippenberg, and Ilgen (2017) have identified three themes that underlie team performance, namely "(a) team tasks and structure; (b) member characteristics and team composition; and (c) team processes and emergent states" (p. 455). This research focuses on one of these emergent states: psychological safety (Mathieu et al., 2017; Newman, Donohue, & Eva, 2017).

Psychological safety has been defined as "a shared belief held by members of a team that the team is safe for interpersonal risk taking" (Edmondson, 1999, p. 350). Examples of activities that can be interpersonally risky are admitting to and discussing errors, asking for feedback and raising new ideas (Newman et al., 2017; Pearsall & Ellis, 2011). Engaging in these activities has been found to raise team performance, e.g. through mechanisms of team and organizational learning (Edmondson, 1999). In addition, research has found a positive relationship between organizational-level psychological safety and overall firm performance (Baer & Frese, 2003).

Typically, psychological safety has been measured by the means of surveys, for which different measures have been developed (e.g. Edmondson, 1999; Nembhard & Edmondson, 2006). However, the usability of surveys is limited by several constraints, mainly: self-report bias (Donaldson & Grant-Vallon, 2002) and non-response bias (Dooley, 2009b; O'Donovan, Van Dun, & McAuliffe, 2020). Respondents answering surveys are inclined to convey an overly positive picture of the situation due to social desirability (Donaldson & Grant-Vallon, 2002). Additionally, they might have biased self-perception. There can be a gap between behaviour the respondents think they engage(d) in and the actual behaviour (Baumeister, Vohs, & Funder, 2007). Furthermore, sample rates are dependent on the respondents' willingness to fill in the survey. This non-response bias can become an issue when there are structural differences between people responding and people refusing to respond as the sample will not be representative of the actual population (Dooley, 2009b).

All these issues compound when the intention is to use repeated measurements to analyse the development of a concept over time (Kozlowski, 2015). Thus, for a dynamic construct such as psychological safety, these limitations of survey-based research can have a great effect. Recently, however, there has been a slow trend of researchers moving towards

observational research methods to study such social concepts (Baumeister et al., 2007). The use of observational, and even video-based research methods, intends to counteract the limitations above and enable a more dynamic measurement (LeBaron, Jarzabkowski, Pratt, & Fetzer, 2018). Observational research uses coding schemes consisting of numerous observable behaviours, called “codes” (Waller & Kaplan, 2018). Researchers observe participants and collect data on their behaviour according to these codes. Naturally, this type of research methods also has inherent flaws, e.g. the reliance on the subjective perception of the researcher (Foster, 2006). Therefore, a combination of observational research methods with more traditional methods is the most viable (Klonek, Gerpott, Lehmann-Willenbrock, & Parker, 2019).

An observational coding scheme for psychological safety has been developed by O'Donovan et al. (2020). The observation scheme is grounded in the general team literature and has been further developed in collaboration with healthcare professionals to address the specific environment of the health care sector as psychological safety is supposed to have particular value in this context (Newman et al., 2017). However, the observation scheme might also be practicable in other sectors, as only few codes are directly related to the health care context.

The current research uses this observation scheme, in addition to traditional survey-based measures, to analyse psychological safety in four studies conducted with work teams in different industry sectors. Doing this, the research intends to answer the following research question:

#### **WHAT IS THE ADDED VALUE OF USING OBSERVATIONAL RESEARCH METHODS IN SURVEY-BASED TEAM PSYCHOLOGICAL SAFETY RESEARCH IN WORK TEAMS?**

In answering these research questions three assumptions are addressed, these being:

- (1) The observable psychological safety related behaviour differs between teams depending on their level of psychological safety.
- (2) Video-based research technology, i.e. The Observer XT, can aid in reliably identifying when behaviours related to psychological safety occur and can enrich data collection.
- (3) Teams with higher psychological safety have higher survey-reported team performance.

Exploring this research question and the assumptions can have incremental theoretical and practical relevance.

On a theoretical level, this research advances group-level psychological safety research. According to Frazier, Fainshmidt, Klinger, Pzeshkan, and Vracheva (2017), who conducted a meta-analysis on psychological safety, empirical research on group-level psychological safety is still scarce, limiting the ability to draw valid conclusions. Moreover, the analysis of behaviour of teams can provide in-depth insights that can reveal underlying

mechanisms which create certain levels of team psychological safety. This can enable researchers to develop more accurate theories on the existence of psychological safety and more specifically on ways to develop psychological safety in teams. Additionally, a validated observation scheme for psychological safety will enhance the reliability of psychological safety measurement. Not only because well-developed observational research methods should provide a more reliable picture but also because using them in combination with other research methods enables better triangulation of results. A recent literature review has also called for the development of alternative methodologies for studying psychological safety (Newman et al., 2017).

On a practical level, an exploration of the usability and value of the psychological safety observational scheme could enable practitioners, such as team leaders, managers, or consultants, to use these schemes to assess psychological safety in their own work environment (O'Donovan et al., 2020). Furthermore, results on specific behaviours that are positively related to psychological safety could inform what behaviours should be stimulated during psychological safety interventions. Lastly, also the measurement of the effectiveness of these interventions could be improved and made more practicable by adding an observational element to the research. This is also what O'Donovan and McAuliffe (2020a) call for in their systematic review of psychological safety interventions.

## THEORETICAL FRAMEWORK

The theoretical framework discusses the conceptual constructs used in this research. First, the central concept of psychological safety is elaborated on, including the closely related concepts of voice and silence. The literature on team performance is reviewed to establish the practical relevance of researching psychological safety. Lastly, the theoretical background of using observational research methods in addition to traditional research methods is discussed.

## PSYCHOLOGICAL SAFETY

Psychological safety describes an environment in which people feel safe to express themselves, e.g. raising (personal) issues and ideas, admitting to error, asking risky questions (Newman et al., 2017; Pearsall & Ellis, 2011), and do not fear that this will have negative consequences for them (Kahn, 1990). Research has termed these actions to express oneself “interpersonal risk taking” (Edmondson, 1999, p. 350).

There have been significant differences found in psychological safety between teams from the same organization (Edmondson, 1999). While employees within one team perceive psychological safety levels similarly, perceptions of employees from other teams can differ. Employees need to take interpersonal risks within their team to align their perspectives and collaborate effectively to reach their shared goals, which can explain why team members have a shared perception of psychological safety within their own team (Edmondson & Lei,

2014). Following this reasoning, it is advocated to consider psychological safety on the team-level (Edmondson & Lei, 2014), which this research does.

Previous research has found various antecedents and outcomes of psychological safety. Antecedents and outcomes exist at three different levels: the individual level, the team level and the organizational level (see Table 1). The outcomes that are studied in this research are marked by bold lettering: voice behaviour and silence behaviour (indirectly), and team performance. It has been chosen to consider the literature on Voice and Silence as these are integral elements of the observation scheme used in this research. The following sections elaborate on these concepts and their relationship with psychological safety.

	Antecedents	Outcomes
<b>Individual-level</b>	Work design: autonomy, role clarity and independence (Frazier et al., 2017)	Task performance (Frazier et al., 2017)
	Help and coaching by team leader coaching (Edmondson, 1999; Newman et al., 2017)	Engagement in quality improvement work (Newman et al., 2017)
		Reduction in errors (Newman et al., 2017)
		Higher satisfaction (Frazier et al., 2017)
		Higher creativity (Frazier et al., 2017)
		Higher work engagement (Frazier et al., 2017)
		<b>Higher voice behaviour</b> (Detert & Burris, 2007; Liang, Farh, & Farh, 2012; Walumbwa & Schaubroeck, 2009)
		<b>Lower silence behaviour</b> (Brinsfield, 2013; Sherf, Parke, & Isaakyan, 2020)
<b>Team-level</b>	Help and coaching by team leader (Edmondson, 1999; Newman et al., 2017)	<b>Team performance</b> (Kostopoulos & Bozionelos, 2011; Newman et al., 2017)
	Integrity of the leader (Newman et al., 2017)	Team learning (Edmondson, 1999; Frazier et al., 2017; Kostopoulos & Bozionelos, 2011; Newman et al., 2017)
	Leader inclusiveness (Newman et al., 2017)	Higher information sharing (Frazier et al., 2017)
	Trust in the leader (Newman et al., 2017)	
<b>Organizational-level</b>	Context support (Edmondson, 1999; Frazier et al., 2017)	Higher firm performance (Baer & Frese, 2003; Edmondson & Lei, 2014)

Table 1: Antecedents and outcomes of psychological safety on three levels

## VOICE AND PSYCHOLOGICAL SAFETY

*Voice behaviour* typically has been defined as employees speaking up with the goal of igniting positive change regarding work-related issues (LePine & Van Dyne, 1998; Morrison, 2014; Van Dyne, Ang, & Botero, 2003). This includes speaking up with new ideas or suggestions or raising awareness about mistakes or problems that have been encountered. People are more likely to engage in *Voice behaviour* when they perceive the impact of speaking up to be high (Sherf et al., 2020). *Voice behaviour* has been found to be beneficial for organizations, improving their overall performance (Detert, Burris, Harrison, & Martin, 2013; MacKenzie, Podsakoff, & Podsakoff, 2011; Nemeth, Connell, Rogers, & Brown, 2006).

However, researchers have conceptualized that voice behaviour does not have to stem from this pro-social intention but can also be grounded in disengagement or self-protection of the speaker (Van Dyne et al., 2003), thus being rather negative. Voice behaviour can then also be categorized as *Acquiescent Voice* or *Defensive Voice* respectively.

People use *Acquiescent Voice* when they are not confident that they can elicit meaningful change (Van Dyne et al., 2003). For example, people disengage from discussions, merely agreeing with what is being said and simply accepting ideas from others, instead of communicating their own opinions or ideas.

*Defensive Voice* can occur when a person is feeling threatened. When engaging in defensive voice, the speaker tries to actively protect themselves from undesired consequences (Van Dyne et al., 2003). Examples of this kind of voice are intentionally diverting attention from a certain issue or blaming others for the issue.

It has been conceptualized that when there is an opportunity for speaking up, i.e. an employee has encountered an issue and is sitting in a meeting with his team, the employee makes a conscious, calculated decision whether to speak up about this issue or not (Detert & Burris, 2007; Liang et al., 2012). This choice is based on the balance between the costs and benefits of speaking up (Liang et al., 2012). Potential costs are negative repercussions from speaking up about sensitive topics, such as ridicule or even negative job consequences, such as limited future job opportunities (Detert & Burris, 2007). Benefits are by large organizational (Klaas, Olson-Buchanan, & Ward, 2012), but there can also be personal benefits, such as admiration or positive job consequences (Detert & Burris, 2007; Morrison, 2014).

Team psychological safety is conceptually related to *Voice behaviour*. When people believe it is safe to take interpersonal risks, the potential costs of speaking up, a form of interpersonal risk-taking, are naturally decreased (Liang et al., 2012). Consequently, the benefits of speaking up exceed the costs, thus making voice the favourable choice (Detert & Burris, 2007). This way psychological safety can be associated with *Voice behaviour*. Empirical studies treating psychological safety as a mediator between different modes of leadership and voice have found a significant positive relationship between psychological safety and voice on their own (Detert & Burris, 2007; Walumbwa & Schaubroeck, 2009).



Liang et al. (2012) also found a significant positive relationship between psychological safety and *Voice behaviour*. More specifically, their research set out to study the causal relationship between several psychological constructs and voice behaviour. Theoretically, voice could not only be an outcome but also an antecedent to psychological safety: It could be that because some people speak up, others interpret that this is appropriate behaviour and that it is safe for themselves to do so in the future (Liang et al., 2012). Over time this could result in a psychologically safe environment. A two-wave panel study showed that there was a significant positive relationship between psychological safety and temporal changes in voice behaviour (Liang et al., 2012). This supports the positioning of voice as an outcome rather than an antecedent of voice.

---

#### SILENCE AND PSYCHOLOGICAL SAFETY

*Silence behaviour*, on the other hand, occurs when a person has an opinion, an idea or a concern, but decides not to voice this (Morrison, 2014). This is inherently different from just being silent, as people can also be silent just because they have nothing to say. The concept of *Silence behaviour*, however, implies that the person has something important to say but purposefully withholds this from their conversational partner(s) (Morrison, 2014). Withholding such information can be inherently detrimental to organizations; constraining organizational change and improvement (Morrison & Milliken, 2000). *Silence behaviour* has been far less researched than *Voice behaviour* even though it can be just as impactful (Morrison & Milliken, 2000; Pinder & Harlos, 2001).

Similar to the three dimensions of *Voice behaviour*, Van Dyne et al. (2003) conceptualized three dimensions of *Silence behaviour*. In the case of *Silence behaviour*, they disagree with the mainstream literature by adding a form of Silence that is Pro-Social, so not detrimental to the society per se. The following three types have been conceptualized: *Defensive*, *Acquiescent* and *Pro-Social Silence*.

*Defensive Silence* comes from fear. People engage in this type of silence when they are afraid of the consequences of voicing their ideas, concerns or opinions. They actively withhold the information in order to protect themselves (Pinder & Harlos, 2001; Van Dyne et al., 2003). Research has found that especially fear of punishment or negative career consequences pushes people to keep silent (Detert & Edmondson, 2011; Milliken, Morrison, & Hewlin, 2003).

*Acquiescent Silence* comes from disengagement. A person who engages in this type of silence, does not want to put in the effort to voice their opinions, ideas or concerns. The person is resigned from the situation or conversation. This can, for instance, be based on the self-belief that the person cannot make meaningful change by speaking up (Van Dyne et al., 2003). Weiss, Kolbe, Grote, Spahn, and Grande (2018) also identify limited self-efficacy as a reason for Silence behaviour. However, recent empirical research identifies perceived impact as only a weak predictor of Silence behaviour (Sherf et al., 2020).

Lastly, *Pro-Social Silence* comes from altruism. When engaging in *Pro-Social Silence* a person actively withholds information because the person thinks that sharing it would be detrimental to the organization. This could, for example, be the case with confidential information or when a person does not complain about circumstances to not burden others (Van Dyne et al., 2003). Not wanting to harm relationship with co-workers has also been identified as a reason why people keep silent, especially in people who highly value interpersonal relationships (Weiss et al., 2014), which could be a form of *Pro-Social Silence* as well.

It has been found that in a psychologically safe environment, people are significantly less inclined to engage in *Silence behaviour* overall (Sherf et al., 2020). This relates to the conceptualization of silence as self-protecting (*Defensive Silence*). When an environment is psychologically safe, interpersonal risks can be taken without fear of implications, therefore self-protection is less relevant and people are not pushed into keeping silent about concerns, ideas or opinions. Interestingly, Sherf et al. (2020) found that psychological safety relates more strongly to *Silence behaviour* than *Voice behaviour*.

Moreover, a climate of fear has been related to higher *Silence behaviour* (Morrison, 2014; Morrison & Milliken, 2000; Pinder & Harlos, 2001). As a psychologically safe environment should diminish fear, this could mean that it elicits less *Silence behaviour*.

Lastly, Brinsfield (2013) has found that psychological safety is negatively correlated with three sub-forms of *Silence behaviour*, amongst which *Defensive Silence*. Thus, higher levels of psychological safety are associated with lower levels of *Silence*.

The literature seems to point towards psychological safety being especially related to *Defensive Silence* rather than *Acquiescent* and *Pro-Social Silence*.

---

#### TEAM PERFORMANCE AND PSYCHOLOGICAL SAFETY

Team performance has varying conceptualizations and is sometimes used interchangeably with the term team effectiveness (e.g. Gibson, Cooper, & Conger, 2009). In this research, exclusively the term team performance will be used. There are two ways to measure the performance of teams, one being the usage of tangible data on team outputs and the other being the usage of perceptions of team members and managers (Mathieu et al., 2017). In this research the perceptions of team performance are assessed in the sense of how effectively the team is working, congruent with the measures used to assess team performance by Gibson et al. (2009).

There are various ways in which psychological safety can enhance team performance. Firstly, psychological safety can influence team performance through other mediators, for example, through team learning (Edmondson, 1999; Kostopoulos & Bozionelos, 2011). Team learning requires employees to generate new ideas and express them openly. A low level of psychological safety, i.e. team members feeling that the risk of embarrassment or critique is high, may obstruct team members' inclination to engage in such behaviour, thus decreasing

the level of team learning and consequently, lowering team performance (Kostopoulos & Bozionelos, 2011).

Secondly, a different perspective sees psychological safety as a moderator of relationships of other constructs with team performance. For example, Martins, Schilpzand, Kirkman, Ivanaj, and Ivanaj (2013) found that psychological safety moderates the relationship between expertness diversity on teams and team performance, where expertness diversity was negatively associated with performance when psychological safety was low, but positively associated with team performance when psychological safety was high. It can be theorized that this effect is due to the team accepting ideas and suggestions of members with differing expertise more easily when there is a climate of psychological safety rather than when there is not.

---

#### DOWNSIDERS OF PSYCHOLOGICAL SAFETY

The elaboration above focuses on the positive sides of psychological safety exclusively but research has found that high levels of psychological safety can have negative impact. Studies have found that psychological safety can be related to unethical behaviour. Pearsall and Ellis (2011) studied the effect of psychological safety on the relationship between two ethical orientations – utilitarianism and formalism – on unethical team behaviour. Utilitarianists make decisions with consideration for the end goals more so than for the means with which they achieve those goals (Brady, 1985). When a decision might violate social norms, a utilitarianist does not see this as a problem as long as the violation is justified by the benefit of achieving the goal. Pearsall and Ellis (2011) found that when teams had utilitarian members and also had high psychological safety, the team was significantly more likely to engage in unethical behaviour than when there was lower psychological safety. Supposedly, this is because the psychologically safe environment enables people with unethical ideas to speak up about them (Pearsall & Ellis, 2011), increasing the likelihood of the ideas being put to use. More recently, in a study of the mediating effect of psychological safety between charismatic leadership and unethical behaviour, a significant direct association between psychological safety on unethical behaviour has been found (Zhang, Liang, Tian, & Tian, 2020).

#### OBSERVATIONAL RESEARCH METHODS

This section presents a rationale for using observational research methods besides traditional ones, such as surveys.

---

#### HISTORY OF OBSERVATIONAL RESEARCH METHODS

Observational research was very common in behavioural studies till the 1980s but since 1986 there has been a steady decline in the usage of this method of data collection (Baumeister et al., 2007). This has been attributed to journals not valuing observational research adequately. Additionally, the failure of finding significant results with observational research is very costly due to the increased effort necessary to conduct observations, compared to, for example, surveys (Baumeister et al., 2007).

Only 2 out of 38 studies in an issue from January 2006 from the *Journal of Personality and Social Psychology* used data derived from studying actual behaviour, i.e. observations (Baumeister et al., 2007). Most studies used self-report measures, particularly questionnaires. This preference for quantitative surveys has also been identified in current organizational behaviour research (Donaldson & Grant-Vallou, 2002), overall team research (Mathieu et al., 2017) and in psychological safety research specifically (Newman et al., 2017). However, researchers are making calls to incorporate alternative methodologies such as observations, to reach a deeper level of understanding of the complexities of psychological safety and its relevance for teams (Edmondson & Lei, 2014; Newman et al., 2017). Recent literature indeed identifies observational research methods to be a slowly emerging, or reappearing, trend in social research (Meinecke, Klonek, & Kauffeld, 2016).

---

#### ISSUES IN SURVEY-BASED RESEARCH

Potential issues that undermine the effectiveness of traditional surveys are associated with self-report bias and non-response bias. Below it is elaborated how these biases affect traditional survey research.

Self-report bias has been conceptualized to surface based on four factors: the true state of affairs, the sensitivity of the researched construct, dispositional characteristics of the respondent and situational characteristics (Donaldson & Grant-Vallou, 2002). An underlying element of these factors is the propensity of respondents to want to convey a positive picture of themselves, called social desirability bias (Baumeister et al., 2007; Donaldson & Grant-Vallou, 2002).

Regarding the true state of affairs, survey respondents have to be able to remember correctly how they felt or what they did in a given situation to answer survey questions truthfully. However, people generally have difficulty remembering and recalling situations, their actions and thoughts in exhaustive detail (LeBaron et al., 2018). The quality of recalled information depends on various factors, such as the time since the event occurred, salience of the event, and also social desirability of the event (Beckett, Da Vanzo, Sastry, Panis, & Peterson, 2001). Incorrect recall of information can lead to a gap between the behaviour that is reported and the behaviour that would actually be observed (Baumeister et al., 2007). Sometimes people are not even aware of their behaviour or factors underlying their behaviour while it occurs which can make the exactness of survey responses even more questionable (Baumeister et al., 2007; Christianson, 2018; Foster, 2006; LeBaron et al., 2018). Therefore, self-report measures, such as surveys, are strongly limited by the subjective perception and remembrance of the respondent at the moment of answering the question (Meinecke et al., 2016).

Additionally, survey questions can create reactivity, in the sense that respondents can feel forced to convey an opinion, feeling or behaviour in their responses only because they are aware of the topic being researched (Hill, White, & Wallace, 2014). This issue could be

aggravated by the aforementioned social desirability bias, potentially biasing the answers of otherwise indifferent participants to favourable responses.

Non-response bias occurs when survey respondents fail to respond to one or several items of the survey or do not complete the survey at all. This can be due to various reasons, e.g. busyness or fear of consequences of the responses (Foster, 2006). The issue with non-response bias is that there may be underlying differences between the group that responded and the group that did not respond (Dooley, 2009b). For example, when studying psychological safety in groups, people who do not feel psychologically safe might not respond to surveys while people feeling safe *do* respond. This results in non-observation bias of the group of people who do not feel psychologically safe and can push the results in a too favourable direction. The benefit of observational research in this sense is that all participants are observed, also participants that might not have participated in the survey, leading to the extraction of a more complete picture of the sampled research subjects (Foster, 2006; O'Donovan et al., 2020).

Lastly, when studying dynamic processes, such as psychological safety, it is advised to conduct longitudinal research in which data is collected at numerous occasions to understand the development of the construct over time (Kozlowski, 2015). However, in this case, the aforementioned issues would compound and pose an even larger constraint: For example, recall bias becomes a bigger issue in longitudinal research, when information is asked about experiences since the last measurement which can be a long time ago (Wang et al., 2017). Moreover, longitudinal survey studies have to deal with decreasing response rates at consecutive data collection waves (Castiglioni, Pforr, & Krieger, 2008; Ployhart & Ward, 2011) due to response exhaustion. Additionally, respondents might remember their previous responses and give the same responses in order to remain consistent.

---

#### HOW CAN THE INCLUSION OF OBSERVATIONAL RESEARCH METHODS PREVENT THIS?

The reliance on subjective perception of participants and their willingness to respond is omitted when using observational research methods as behaviour is assessed for all participants as it occurs in real-time. During observations the whole data collection is subject to the perception of the specific researcher. The knowledge and personal interpretations of the observer could bias the results (Foster, 2006). To account for this, it is advocated for systematic observational research that several researchers observe the same situation (Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000). The observations and coding of the various researchers can then be compared to assess the reliability of the observations through which a degree of objectivity should be achieved.

However, it can be detrimental to have several researchers observing the participants in real life since the presence of researchers can cause reactivity, leading to participants altering their behaviour (Foster, 2006). It can be assumed that this issue intensifies with an increasing number of observers present.

On a different note, it can be difficult for researchers to analyse behaviours while they occur as behaviours can be very short-lived. Moreover, they are embedded in numerous other behaviours that might be irrelevant to the study. Researchers need to be able to identify and separate these behaviours on the spot during real-time observational research (Christianson, 2018; Noldus et al., 2000). This difficulty also, naturally, increases with the number of research subjects to be observed (Meinecke et al., 2016).

To overcome these difficulties, the next step is to conduct observations based on video-recordings of the situations to be studied, as will be explained below.

---

#### VIDEO-BASED OBSERVATIONS

According to Christianson (2018), while the potential of video for research has been widely discussed in social sciences like sociology and communications sciences, it has only recently gained attention from the organizational sciences.

A benefit of video recording observational settings is that videos can be revisited by the researcher (or additional researchers) multiple times to ensure correct and reliable coding (LeBaron et al., 2018; Pugliese, Nicholson, & Bezemer, 2015). More specifically, videos can be paused, rewound, and slowed to capture even more details in the behaviour of participants (Christianson, 2018; Noldus et al., 2000). Such technical features enable micro-coding, an approach with which the precise timing and frequency of behaviour is minutely assessed (Waller & Kaplan, 2018). This can facilitate the analysis of sequences of behaviour such as the effect of one behaviour on behaviour in the coming minutes (LeBaron et al., 2018; Meinecke et al., 2016). Considering the sequence of behaviours is relevant to understanding the meaning of the behaviour. Often a behaviour is given meaning by the behaviours that occurred before it and/or after it (LeBaron et al., 2018). For example, shaking your head would generally imply disagreement. However, when it occurs as a reaction to a negatively formulated statement, shaking your head can mean that you agree with the negatively formulated statement. If one was to look only at the single behaviour of shaking one's head, it would have been interpreted as disagreement which would have been untrue in this example. Analysing these kinds of sequences can uncover which behaviours stimulate or, in contrast, stifle psychological safety related behaviour.

Moreover, studying sequences of behaviour can reveal patterns, i.e. when regularly the same behaviours occur subsequently. There are computer programmes, such as Theme, which are used by researchers to facilitate recognizing such patterns (Waller & Kaplan, 2018).

---

#### CHALLENGES IN VIDEO-BASED OBSERVATION

However, there are also challenges to video-based observation. Video-recording presents the challenge of deciding from which angle the participants will be recorded. This choice can already influence data analysis and even the outcomes of the analysis, so it is critical to the research process (LeBaron et al., 2018). For example, a video camera can be placed amongst the participants and therefore record the situation from the viewpoint of a participant or it

can be placed in a birds-eye view where the whole situation is recorded from an outside perspective (LeBaron et al., 2018). These two perspectives will give the researcher different insights about the participants' behaviour. For example, when filming from the participant-view, chances are that not all participants will be visible on the recording which might impede analysis. Consequently, researchers should deliberately consider the placement of the video camera based on the goal of their research. When using video that has been pre-recorded by other researchers, the researcher at hand should also recognize how the placement of the camera can influence his results.

Additionally, concerns can arise regarding participants' reactivity to video cameras. Indeed, in the medical field, concerns have been voiced that the presence of video camera could alter behaviour of participants (Penner et al., 2007). However, subsequent research has found that, in the medical field, only 0.1% of behaviour during recordings was related to the video camera and when this occurred, it was predominantly in the beginning of the situation recorded (Penner et al., 2007). In a business setting, similarly, research using video-recorded board meetings has found that cameras do not alter the behaviour of participants during the meeting, except for marginally at the very start of the recording (Pugliese et al., 2015). Moreover, when asked, participants of video-based research emphasized that the video cameras did not alter their behaviour and interactions during the meeting (Pugliese et al., 2015). Furthermore, previous video-based research found through surveys that behaviour during recorded meetings was representative of non-recorded meetings (Hoozeboom & Wilderom, 2020). In conclusion, while researchers should keep an eye on behaviour signalling reactivity, overall, it can be said that concerns about the reactivity of video cameras can be neglected for this research and it can be expected that recorded meetings are representative of 'usual' meetings that are not recorded.

However, a problem that can intensify when using videos in research on sensitive topics, such as psychological safety, is the aforementioned non-response bias. As mentioned above, non-response bias can be structural where e.g. only people that feel psychologically safe respond. When asking to video-record a meeting for psychological safety research, teams with low psychological safety might not allow it while teams with high psychological safety do. That way, only highly psychologically safe teams would be observed. This could be remedied by either recording videos for assessment of several less-sensitive concepts next to psychological safety, where teams can gain knowledge about their practices on several constructs which could off-weigh the costs of getting video-recorded. A different approach would be to analyse psychological safety in teams that have already been recorded for other purposes if it is allowed to re-use these videos.

---

## CONCLUSION

The elaboration above shows how observational methods can counteract some of the issues encountered when using self-report research methods, such as surveys but also the challenges of engaging in observational research. Observational methods should not be seen as a replacement but rather as an extension of self-report methods (Meinecke et al., 2016).

In fact, researchers advise to combine observational data with data generated from traditional methods, such as surveys (Klonek et al., 2019). Using a mixed methods approach should not only allow for triangulation of results but, additionally, enable a more detailed understanding of the phenomena with the potential to discover new phenomena.

## METHODOLOGY

### OVERALL RESEARCH DESIGN

This research used three different samples from Dutch organisations that have been collected in previous studies. In this research, a mixed-method approach of observations and surveys has been used. While both measures were evaluated quantitatively, the observations were also analysed qualitatively. This combination of quantitative and qualitative analysis is proposed by Edmondson and McManus (2007) in situations where recently developed measures are used or underlying mechanisms are analysed. Both of these come forth in this research as the observational scheme that was used has been piloted only recently and the analysis of observations was used to detect differences in specific behaviours that underlie psychological safety. While the observations were used to triangulate the survey findings, i.e. to assess whether the same results are found when employing different methods (Tracy, 2010), the observations were also used to crystallize the findings, i.e. to get additional insights and get an in-depth understanding of the concept (Tracy, 2010).

The observational scheme that is used throughout the research has recently been developed by O'Donovan et al. (2020) using research by Hoenderdos (2013) as the foundation.

During all observational analysis a static approach was taken, meaning that the differences in behaviour across teams were analysed rather than the differences in behaviour within one team over time (Klonek et al., 2019).

During all quantitative analyses, non-normality of the data is assumed and a minimum significance level of 0.2 is used. The reasoning for these choices is explained in Appendix I and II.

Table 2 gives an overview of the pilot study and four studies in this research. The design, sampling, methods and analysis of each study is further elaborated below.



Study	Constructs	Method	Method source	Items	Example item	N =	Respondents
<b>Pilot study</b>	Team psychological safety	Observation	O'Donovan et al. (2020)	62	"Denying fault or blame other"	2	Team members
<b>Study 1</b>	Team psychological safety	Survey	Edmondson (1999)	6	"It is safe to take risks within this team."	4	Team members; Team leaders
		Observation	O'Donovan et al. (2020) (adapted)	158	"Denying fault or blame other"		Team members; Team leaders
	Team performance	Survey	Van Den Bossche, Gijssels, Segers, and Kirschner (2006)	4	"We have completed the task in a way we all agree upon."		Team members; Team leaders
<b>Study 2</b>	Team psychological safety	Survey	Nembhard and Edmondson (2006)	4	"If you make a mistake in this team, it tends to be held against you."	7	Team members
		Observation	O'Donovan et al. (2020) (adapted)	158	"Denying fault or blame other"		Team members; Team leaders
	Team performance	Survey	Gibson et al. (2009)	4	"This team is consistently a high performing team."		Team leaders
<b>Study 3</b>	Team psychological safety	Survey	Nembhard and Edmondson (2006)	4	"If you make a mistake in this team, it tends to be held against you."	6	Team members
		Observation	O'Donovan et al. (2020) (adapted)	63	"Denying fault or blame other"		Team members
	Team performance	Survey	Gibson et al. (2009)	4	"This team is consistently a high performing team."		Team members; Other related employees
<b>Study 4</b>	Team psychological safety	Observation	O'Donovan et al. (2020) (adapted)	63	"Denying fault or blame other"	1	Team members; Team leader

Table 2: Overview of studies

## PILOT STUDY

### DESIGN

Before any of the other studies were conducted, a pilot study was done, in which the researcher acquainted herself with the observation scheme, tried out both naked-eye and computer-aided observing, and adapted the observation scheme based on experience. The computer-aided observing during this study was also conducted by an additional observer, who also contributed to adapting the observation scheme.

All adaptations were discussed with one of the researchers who developed the original observation scheme.

The video in The Observer XT was coded in three increments: 2 of 10 minutes and 1 of 5 minutes. Between each coding session the two observers corresponded on their experiences and where necessary adapted the observation scheme. Possible adaptations included refining definitions of behaviours and re-formulating behaviours themselves, as well as omitting and including new behaviours.

### SAMPLING

For the pilot study, two agile squads from a large Dutch organization were selected that have been video-recorded for the purpose of other research. The videos show retrospective meetings, in which the squads discussed what went well during their sprint and what could be improved in the future (Annosi, Magnusson, Martini, & Appio, 2016). The videos were selected based on the quality and angle of recordings.

### METHODS

The observation scheme developed by O'Donovan et al. (2020) was used throughout the pilot. This scale consists of a total of 31 behaviours in seven behavioural categories that have been categorized to be indicative of high or low psychological safety. Behaviours on the observation scheme being indicative of high psychological safety were *Voice behaviours*, *Supportive behaviours*, *Learning or improvement-oriented behaviours*, and *Familiarity behaviours*. Behaviours on the observation scheme being indicative of low psychological safety were *Defensive voice behaviours*, *Silence behaviours* and *Unsupportive behaviours*. An example of a behaviour that could be coded is *Denying fault or blame others*.

The observation scheme allowed for coding of behaviour in five directions: how team members interact with the team leader (TL/TM), how the team leader interacts with the team members (TM/TL), how individual team members interact with each other (TM/TM), how the team leader interacts with the team as a whole (Team/TL) and how team members interact with the team as a whole (Team/TM). However, the agile squads in this study followed the Scrum Methodology, meaning that they were self-managing and did not have a team leader (Annosi et al., 2016). So, the directions pertaining to the team leader were not used during

this pilot study. The total number of items that could be coded in the pilot study was thus 62. The original observation scheme can be found in Appendix I.

Both squads were observed with the naked-eye by one researcher. Only the second squad was, additionally, observed in The Observer XT by two researchers. Before all naked-eye observations, the researcher read the transcript of the meeting to get acquainted to its content.

---

#### ANALYSIS

The qualitative experience the researchers got through testing the observation scheme informed the adaption of the scheme.

For the increments that were coded in The Observer XT inter-rater reliability scores were calculated to see initial agreement and monitor whether agreement increased after each adaptation of the observation scheme.

---

### STUDY 1

---

#### DESIGN

This study used a mixed-methods approach to data collection using surveys and observations of work teams. The constructs of team psychological safety and team performance were analysed. Both constructs were assessed through surveys and both team members and team leaders were surveyed. The surveys were filled in before the recorded meetings. Team psychological safety was, additionally, assessed through observations. Observations were conducted on the basis of video-recorded team meetings.

---

#### SAMPLING

In this study, data was collected at four lean teams. For all constructs, team members as well as team leaders were surveyed. The sample included 54 individuals, of which 26 were male and 28 were female. The average age of the respondents was 47 years, with a range from 19 to 62 years old.

Regarding the video, the recordings of one team only showed a single person of the team and this team, therefore, was excluded from the research. The final sample consisted of 4 teams.

The recorded meetings were three daily stand-ups and one weekly progress meeting. Daily stand-up meetings intend to discuss what members have done since the last stand-up, what team members are planning to do until the next stand-up and what issues could hinder the completion of these tasks (Stray, Sjøberg, & Dybå, 2016). During the weekly progress meeting, the team's performance is discussed. Accordingly, the length of the recordings varied, ranging from 2 minutes to 12 minutes. The average length was 7 minutes. Also, the number of team members varied from 5 to 8 people. All but one respondent were Dutch.

Survey data was collected on more team members than were present at the observed meeting for several teams. However, due to anonymization of the data, it was not possible to only select the data from the observed participants during analysis.

---

## METHODS

The survey included items on team psychological safety and team performance.

Team psychological safety was measured with a survey based on Edmondson (1999). Six of the items from this scale were used. An example item from this scale is “It is safe to take risks within this team”. All items were translated into Dutch. Items 1 and 2 were reverse coded for analysis. Cronbach’s alpha for this scale was 0.613 with 6 items. When deleting item 2 a Cronbach’s alpha of 0.627 was achieved. Deleting more variables did not yield a higher Cronbach’s alpha, so the 0.627 had to be accepted. This means that the scale was not fully reliable. Because the individual responses had to be aggregated to the team-level, inter-rater agreement was checked using the rwg (LeBreton & Senter, 2008). This measure has to be at least 0.8 to allow for aggregation. Rwg for team psychological safety was higher than 0.8 in all teams, thus, the individual responses could be aggregated.

Team performance was measured based on Van Den Bossche et al. (2006). Their measure for team performance consists of four items, e.g. “We have completed the task in a way we all agree upon”. All items were translated into Dutch. Cronbach’s alpha for this scale was 0.753. Rwg for team performance was also higher than 0.8 in all teams, so the individual responses could be aggregated to the team-level.

Additionally, the video recordings were assessed for team psychological safety using the adapted observation scheme based on O'Donovan et al. (2020). In total the observation scheme for this study encompassed 158 distinct items due to the 5 different levels on which behaviour was observed.

At all meetings a researcher was present to record the meeting. All meetings were recorded with two cameras, one focusing on the team leader and one focusing on the team members. For both recordings a participant view was chosen. In Teams 1, 2 and 4 the camera recording the team members was mobile, meaning that it varied which team members were visible. The number of team members that were observed was adapted based on how much of each team member was visible.

Before coding the videos, the researcher read the transcript of the meeting to get to know its setting and content, allowing for smoother coding. The videos were coded in one go, counting the number of times the behaviours from the coding scheme could be observed. This was done in one sitting to ensure that each video was watched in similar detail and the results could be compared. The researcher was, thus, not allowed to jump back and forth within the video.

To measure some reliability in coding the total number of observed behaviours per 10 minutes and 5 people was compared across teams. Additionally, the relationship with the PS ratio was calculated to check whether the total number of observations influenced the

observed level of psychological safety. The PS ratio was an indication of the level of psychological safety observed and was calculated by dividing the number of behaviours observed that are indicative of high psychological safety by the number of behaviours observed that are indicative of low psychological safety. Consequently, a higher PS ratio indicates lower psychological safety and vice versa.

The total number of observations ranged from 74.6 to 105.8. Notably, for three teams the total observed behaviour was very similar. Only for Team 2 the total observed behaviour was much higher.

Checking for the relationship between the total observed behaviour and the PS ratio, no significant relationship was found ( $r = -.40$ ;  $p = .60$ ). Thus, observing more behaviour did not impact the observed level of psychological safety.

---

#### ANALYSIS

Study 1 was analysed with three goals in mind (1) exploring the relationship between observed team psychological safety and survey-measured team psychological safety, (2) exploring how specific behaviours and behavioural categories relate to survey-measured team psychological safety and (3) exploring the relevance of team psychological safety in association with team performance.

Due to the particularly small sample size statistical analysis is very unreliable. Therefore, only general statistical correlations were assessed: PS ratio and surveyed team psychological safety, surveyed team psychological safety and behavioural categories in all directions, and surveyed team psychological safety and specific behaviours in all directions. Before analysis, all observational data was averaged to depict the behaviour that would be seen when observing 5 people for 10 minutes.

Finally, for each team individually, qualitative observations were compared to quantitative key findings to crystallize the results. Key findings included the surveyed level of team psychological safety, the PS ratio, and the five most observed behaviours.

---

#### STUDY 2

---

##### DESIGN

Study 2 followed the same design as Study 1. However, in this study team psychological safety was only surveyed with team members and team performance was only surveyed with team leaders.

---

##### SAMPLING

Potential organisations that were adopting lean practices and continuous improvement for at least a year and had shown interest in previous studies of dr. Van Dun were contacted about the research. Only operating level teams were sampled.

In total, 185 companies were invited, of which 85 companies responded. The researchers found 33 of these companies to be fitting the research goals and invited these for a follow-up phone call.

The phone call resulted in 23 companies not being fit for the research, leaving 10 companies. The 10 teams from these companies have 14 team leaders and 96 individual team members. The 10 teams were from various industries, being healthcare, services, production, retirement, human resource and the Dutch ministry of justice and security.

From 7 of these teams, a meeting was taped at which 4-10 team members participated. As the research intended to relate video-based analysis with surveys, only these 7 teams were included in this study.

Two types of meetings were recorded: For Team 1, 2, 3, 4, and 6 daily stand-up meetings were recorded. For Team 5 and 7 weekly stand-up meetings were recorded. Weekly stand-up meetings are structured in the same way as daily stand-up meetings but occur only once a week (Verhelst, n.d.). The length of the videos ranged from eight minutes to almost forty minutes. The average length of the videos was 19 minutes.

58 survey respondents were recruited of which 9 team leaders and 49 team members. 34 of them were male and 12 are female, for 12 people information on gender was missing. The average age of respondents was 42 years with a range from 23 to 63 years. Nationality was not surveyed but as the survey was conducted in Dutch, it can be assumed that the majority of respondents is Dutch. Similar to study 1, survey data was collected on more team members than were present at the observed meeting but due to anonymization it was not possible to match the data. Therefore, all responses are used.

---

## METHODS

The survey included items on team psychological safety and team performance.

Team psychological safety was measured with 4 items developed by Nembhard and Edmondson (2006). This scale is a shortened version of the survey developed by Edmondson (1999). An example item of the scale used in this study is “If you make a mistake in this team, it tends to be held against you”. All questions were translated into Dutch. Item 1 was reverse coded for analysis. Cronbach’s alpha for this scale was 0.801. Rwg was above 0.8 for all teams but one. This team had a rwg of 0.76. The data were still aggregated to the team-level but when analysing Team 2 the lower rwg had to be kept in mind.

Team performance was measured with 4 items developed by Gibson et al. (2009). Example items were “This team makes few mistakes” and “This team is consistently a high performing team.” All questions were translated into Dutch. Cronbach’s alpha was 0.667. Deleting item 4 yielded a Cronbach’s alpha of 0.737. So, item 4 was deleted. Data on team performance was missing for Team 7.

During all meetings, one or two researchers were present to record it. All recordings were made from the participant-view. Cameras were stable, meaning they record the same frame throughout the whole meeting.

Additionally, the video recordings were assessed for team psychological safety using the adapted observation scheme based on O'Donovan et al. (2020). Items and method of observation were elaborated on in the Pilot Study and Study 1. The total standardized number of observations per team in this study ranged from 44.14 to 144.4. This is a very large spread. It could be explained by different meeting styles and paces. Looking at the correlation with the PS ratio, a moderately significant relationship was found ( $r = -.679$ ;  $p = .094$ ). The negative direction of this correlation indicates that teams in which more behaviour was observed, structurally had a lower PS ratio. So, observing more behaviours was related to higher psychological safety. This had to be kept in mind during analysis as it could explain why certain differences between teams were found.

---

## ANALYSIS

Study 2 was analysed with three goals in mind (1) exploring associations between observed team psychological safety and survey-measured team psychological safety, (2) exploring associations between observed psychological safety-related behaviour and survey-measured team psychological safety and (3) exploring the relevance of team psychological safety in association with team performance. All quantitative statistical analyses were conducted using SPSS software.

Analysis followed these steps:

- (1) The observational analysis of the videos was conducted, using three steps: (a) conducting the observations of the videos selected for this research; (b) standardizing the counts to make up for differing video lengths and differing number of participants (Meinecke et al., 2016). The new values showed how often behaviours were observed when watching a 10-minute long video of five people; (c) calculating the counts of behaviours per behavioural category, the counts for behaviours that relate to high psychological safety and behaviours that relate to low psychological safety based on classifications developed by O'Donovan et al. (2020).
- (2) Calculating the PS ratio. This ratio shows how much behaviour related to low psychological safety was observed in comparison to behaviour related to high psychological safety. The closer the PS ratio is to zero, the higher the observed psychological safety in the team.
- (3) Correlating the different behavioural categories and specific behaviours with surveyed team psychological safety scores on all five levels, as well as when counts from all five levels are summed. Regarding the specific behaviours, only behaviours that were observed at least ten times were considered. Findings on behaviours that were observed less than ten times could be coincidental rather than structural.

- (4) Correlating surveyed team performance with surveyed team psychological safety to look into the relevance of psychological safety.
- (5) Qualitatively, assessing the top five behaviours observed in each team.
- (6) Qualitatively assessing what other tendencies the team had.

## STUDY 3

### DESIGN

This study analysed agile squads from a large organization in the Netherlands. Due to the agile terminology “team(s)” in this study were consistently replaced by “squad(s)”. The data were collected via surveys and observations. The squads were visited at three different time points within one sprint: for the first meeting of the sprint where the sprint was started up, for the second meeting of the sprint where the progress and performance so far was discussed, and for the last meeting of the sprint which was a retrospective on the squads’ achievements and collaborations in the finished sprint. This retrospective meeting was used for the observational analysis as it provided an interesting context for analysing psychological safety due to the focus on voicing what went well and what did not. Surveys were conducted at the second and third meeting. At the second meeting individual and team psychological safety was assessed and at the third meeting team performance and, again, individual psychological safety was assessed.

### SAMPLING

This research sampled 9 agile squads. However, for 2 squads survey data on psychological safety was missing and for one squad the third meeting was not recorded. Therefore, the final sample size for this study was 6 squads. In total, 38 people responded to the survey, of which 12 females and 25 males. The average age of respondents was 36 years with a range from 22 to 59 years. 22 of the respondents were Dutch, one was Belgian, six were English, one was Polish, two were Spanish and five belonged to the category ‘Other’.

The recorded meetings were retrospectives, which were meetings in which the squad reviewed their past sprint performance and came up with improvement points for future sprints (Annosi et al., 2016). The length of the meetings ranged from 34 minutes to 1 hour and 43 minutes. The average length was 58 minutes.

While psychological safety was surveyed only with team members, team performance was surveyed with so-called “experts” as well. These were the agile coach, product area lead and tribe lead. The agile coach was actually part of the squad, while the product area lead related to several squads. The tribe lead was positioned higher than the product area lead and related to even more squads.

In this study, it was possible to select only the survey responses of the team members present during the observed meeting.



---

## METHODS

Individual psychological safety was measured using 3 items developed by Detert and Burris (2007). An example item is “It is safe for me to speak up around here”. Cronbach’s alpha for this scale was 0.92 for responses from the second meeting and 0.957 for responses from the third meeting.

Squad psychological safety was measured using 4 items that have been developed by Nembhard and Edmondson (2006). This scale is a shortened version of the survey developed by Edmondson (1999). An example item of the scale used in this study is “If you make a mistake in this team, it tends to be held against you”. Item 1 had to be reverse coded for analysis. Cronbach’s alpha was 0.604. This is lower than the acceptable reliability level of 0.7. Omitting items did not yield a higher Cronbach’s alpha, so this level had to be accepted. Consequently, the reliability of squad psychological safety for this study was delimited. Rwg was above 0.8 for all squads indicating sufficient agreement between squad members to aggregate individual responses to the team-level.

Squad performance was measured using 4 items developed by Gibson et al. (2009) at the last meeting. This measure fits the research as it relates to the productivity and efficiency of the team, as can be seen by its items of “This team makes few mistakes” and “This team is consistently a high performing team.” Cronbach’s alpha was 0.767. Rwg for squad-rated as well as expert-rated squad performance was above 0.8 for all squads indicating that both measures could be aggregated to the team-level.

For each squad, one recording of a squad meeting was assessed. During the meeting no researcher was present in the room and all meetings were recorded from a bird’s eye view. The videos were coded using the adapted observation scheme based on O’Donovan et al. (2020). However, as there are no team leaders in agile squads, the levels pertaining to a team leader were omitted. This resulted in a total of 63 distinct items. The total standardized number of observations per squad in this study ranged from 57.61 to 97.82, resembling a moderate spread and indicating relative consistency in the coding of the observer across squads. The correlation between total behaviours observed and the PS ratio approached marginal significance ( $r = .600$ ;  $p = .208$ ). The direction of the relationship indicated that teams in which more behaviour was observed, could structurally have higher PS ratios, thus lower psychological safety. This had to be kept in mind when evaluating the results of the study.

---

## ANALYSIS

The analysis of Study 3 followed the procedure of Study 2 exactly with the exception that only behaviour between individual team members (TM/TM) and between team members and the team as a whole (Team/TM) were considered. Moreover, since in this study it was possible to match the observed participants to their survey responses, only these survey responses were considered.

## STUDY 4

### DESIGN

The intention of the last study was to explore the possibilities of using the observational scheme of O'Donovan et al. (2020) for analysis of recorded team meetings in the computer program “The Observer”.

### SAMPLING

For this study one team from Study 3 was sampled. It was selected based on its meeting length as coding in The Observer is a time-consuming activity. The length of this video was 36 minutes. It was the meeting of squad 3.

### METHODS

The video was loaded into the computer programme The Observer XT. The adapted observational scheme based on O'Donovan et al. (2020) was used to analyse the videos with this computer programme. Behaviours from the observational scheme were assigned to specific minutes and second in the video when the behaviours started and stopped. This means that the data included not only counts of occurred behaviours but also the duration of occurred behaviours. Additionally, it was coded which team members were engaging in each behaviour.

Two observers separately coded the video and then met to discuss their codes and make a “golden file” which should display the true behaviour. Due to time constraints only for the first 20 minutes of the video a golden file was made. Nevertheless, the separate coding was done for the whole video.

### ANALYSIS

The researcher engaged in quantitative and qualitative analysis of the behaviours identified in the video.

Quantitatively, the PS ratio resulting from the coding in this study was compared to the PS ratio from the naked eye observation of the video.

Qualitatively, the top five scored behaviours from the coding in The Observer XT were compared to the top five score behaviours from the naked eye observation. This was also compared to the five behaviours that had the longest duration in the computer-aided coding. Moreover, the researcher elaborated on the qualitative experience of naked eye coding versus coding in The Observer XT.

## FINDINGS

## PILOT STUDY

## NAKED-EYE OBSERVATIONS

Testing the observation scheme with the naked eye observation was mainly done to get the researcher acquainted to the coding scheme and to observing behaviour in general. However, there were also a few adaptations made to the scheme after the naked eye observations. All adaptations can be found in Appendix IV.

## COMPUTER-AIDED OBSERVATIONS

Table 3 shows how agreement between the two observers developed during the three coding sessions. It can be seen that after each round the agreement increased. This can be attributed to both, the researchers getting more familiar with the codes themselves, and the researchers adapting the codebook to make it more explicit.

	Video length	Agreement	Kappa
Round 1	10 min	9.09%	-.04
Round 2	10 min	16,37%	0.16
Round 3	5 min	26,46%	0.2

Table 3: Inter-rater agreement during three rounds of testing the observation scheme

However, even after three rounds kappa was only 0.2 which is nowhere near the 0.7 that is advised by literature (Waller & Kaplan, 2018). More test rounds could have further improved agreement but this was not possible within the time frame of this study. Moreover, the observation scheme was still quite complex leading to a lot of inconsistencies in interpretation.

In Appendix IV the changes that were made to the observation scheme after each round are summarized. In Appendix V a checklist is presented that has been made after round two that should be followed when observing team meetings in The Observer using the Psychological Safety Observation Scheme. The final scheme includes 35 behaviours and can be found in Appendix VI.

## STUDY 1

## RELATIONSHIP OBSERVED PS RATIOS AND TPS

Table 4 shows the correlations of the PS ratios with team psychological safety. Only the relationship of PS ratio of all behaviours combined with team psychological safety is marginally significant. Also, this relationship follows an unexpected direction. The positive relationship indicates that a higher PS ratio relates to higher survey measured psychological safety. Theoretically, a lower PS ratio should indicate higher psychological safety.

Furthermore, the results for the PS ratio on the TM/TM level and the Team/TM level are striking. They indicate that these ratios correlate 100% with TPS. And, again, this correlation

would be in the unexpected direction, meaning that observing more behaviour that should indicate lower psychological safety, actually is related to higher psychological safety. These interpretations should, however, be considered very cautiously as the sample size in this study is extremely small ( $n = 4$ ). The small sample size means that it could be coincidental rather than systemic that the variables relate to each other.

	M	SD	TPS
TPS	4.643	.492	
PS ratio (TL/TM)	.025	.050	.258
PS ratio (TM/TL)	1.172	1.895	.600
PS ratio (TM/TM)	.449	.188	1
PS ratio (Team/TL)	.023	.045	-.775
PS ratio (Team/TM)	.826	1.023	1
PS ratio (all)	.355	.342	.800*

Table 4: Spearman's Rho correlations TPS with PS ratios

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP ALL OBSERVED BEHAVIOURAL CATEGORIES AND TPS

Table 5 shows the relationship of the observed behavioural categories from all levels combined with team psychological safety. The relationship between *Unsupportive behaviour* and team psychological safety is marginally significant. The association is positive which is unexpected. Further analysis will look into specific behaviours of this category to find an explanation for this correlation.

	M	SD	TPS
TPS	4.643	.492	
Voice behaviour	27.194		-.400
Defensive voice behaviour	.590		.316
Silence behaviour	7.498		.600
Supportive behaviour	22.674		-.600
Unsupportive behaviour	10.430		.800*
Learning and improvement behaviour	11.656		-1
Familiarity behaviour	3.897		.738

Table 5: Spearman's Rho correlations TPS with behavioural categories from all levels combined

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP SPECIFIC BEHAVIOURS OBSERVED IN ALL DIRECTIONS COMBINED AND TPS

Table 6 shows which specific behaviours are correlated with team psychological safety when combining all directions that can be measured in one score. All relationships are in the unexpected direction, giving support for the inadequacy of statistical analysis with a sample size as small as 4.

	M	SD	TPS
<b>TPS</b>	4.643	.492	
<b>Disagreeing</b>	2.574	3.921	-.949**
<b>Sharing procedures, knowledge and experience</b>	1.298	1.741	-.949**
<b>Interrupting</b>	2.704	2.167	-.800*
<b>Discussion within small sub-groups</b>	4.188	4.942	.800*
<b>Verifying progress and performance</b>	1.172	1.738	-.949**
<b>Asking for feedback</b>	.586	.869	-.949**
<b>Informing about issues and mistakes</b>	3.380	1.483	-.800*
<b>Speaking up with ideas</b>	2.032	2.387	-.949**

Table 6: Spearman's Rho correlations TPS with specific behaviours from all levels combined

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

### CONCLUSION OBSERVED BEHAVIOURS AND TPS

As mentioned above, it is difficult to infer statistical relationships from the data in this study, as the sample size is so small. The data do show that there might be a relationship between the observed level of psychological safety (the PS ratio) and the surveyed level of psychological safety. However, this relationship is in the unexpected direction. The analysis of specific behaviours shows that all relevant correlations are in the unexpected direction, potentially explaining why also the PS ratio is related to psychological safety in the unexpected direction.

### TEAM PERFORMANCE AND TEAM PSYCHOLOGICAL SAFETY

The relationship between team performance and team psychological safety is marginally significant ( $r = .800$ ;  $p = .200$ ). This indicates that teams that have high psychological safety would also have high team performance.

### TRIANGULATING THE TEAM-SPECIFIC RESULTS

Next, key findings for each team separately are considered and compared to qualitative observations of the teams. Table 7 shows the key quantitative findings.

	TPS	PS ratio	5 most observed behaviours
<b>Team 1</b>	5.11	0.32	1. TL/TM Active listening 1. Team/TL Providing information 3. TL/TM Asking for further clarification 4. TM/TM Providing negative feedback (constructively) 5. TM/TM Discussions within small subgroups 5. TM/TM Reacting cold/ignoring a joke 5. TM/TM Making or laughing about a joke

<b>Team 2</b>	4.00	0.11	<ol style="list-style-type: none"> <li>1. Team/TL Providing information</li> <li>1. Team/TM Providing information</li> <li>3. Team/TL Asking for further clarification</li> <li>4. TL/TM Agreeing</li> <li>5. TM/TL Disagreeing</li> <li>5. Team/TM Asking for further clarification</li> </ol>
<b>Team 3</b>	4.53	0.15	<ol style="list-style-type: none"> <li>1. TL/TM Active listening</li> <li>2. TM/TM Active listening</li> <li>3. TM/TL Agreeing</li> <li>4. Team/TM Closed body language</li> <li>5. TM/TM Interrupting</li> <li>5. Team/TM Providing information</li> </ol>
<b>Team 4</b>	4.93	0.85	<ol style="list-style-type: none"> <li>1. Team/TM Closed body language</li> <li>1. TM/TM Discussions within small sub-groups</li> <li>3. Team/TL Asking for ideas, help or solutions</li> <li>4. TL/TM Reacting cold/ignoring a joke</li> <li>5. TL/TM Active listening</li> <li>5. TL/TM Making or laughing about a joke</li> <li>5. Team/TL Providing information</li> </ol>

*Table 7: Overview key figures for each team*

#### TEAM 1

This team has a high surveyed team psychological safety score and also a relatively high PS ratio, which contradicts each other. Also, two of the top five most observed behaviours are behaviours that relate to low psychological safety. The discrepancy between the survey data and the quantitative observational data could be explained by the qualitative observations: From this team, a very short meeting was recorded in which the team leader shortly addressed the team and then handed out papers with the results of the week. The rest of the meeting team members were individually looking at the sheets and sometimes making comments to each other on the information which was coded as discussion within small sub-groups.

In conclusion, the observation and the survey results do not align. This means that the results of this team would indicate that the observatory research method is not fit for use in concert with the survey.

#### TEAM 2

This team has neither high nor low psychological safety according to the survey. The observations, however, show a very low PS ratio, indicating high psychological safety. What can be seen from the top five observed behaviours is that four of the behaviours are in relation to the team leader. This was also notable in the qualitative observation. For a major part of the meeting the team leader was talking. One team member was talking for most of

the rest of the time and the other team members were rather quiet with only one or two people sometimes giving their opinions. However, they were listening and not engaging in any disengagement behaviour per se.

In conclusion, the observations of this team do not fully align with the survey measured psychological safety.

---

#### TEAM 3

This team has moderate levels of psychological safety according to the survey and a rather low PS ratio. The top five behaviours show two behaviours that are, theoretically, indicative of low psychological safety: *Team/TM Closed body language* and *TM/TM Interrupting*.

Qualitatively, in this meeting team members were actively raising issues, discussing about them and planning how they can solve them. During these discussions, the behaviour of interrupting one another was visible a lot, which is implicated by the inclusion of this behaviour in the top five most observed behaviours. This behaviour would indicate a high level of psychological safety which makes it unexpected that the survey measured psychological safety is rather low.

Concluding, in this team the results from the observations do not align with the survey measured psychological safety.

---

#### TEAM 4

This team has a relatively high survey measured level of psychological safety but also a very high PS ratio. The top five most observed behaviours also include three negative behaviours: *Team/TM Closed body language*, *TM/TM Discussions within small sub-groups*, and *TL/TM Reacting cold/ignoring a joke*.

From the qualitative observations, it can be said that it was an extremely short meeting with a lot of participants. The team leader was the only one speaking, with the exception of a short question by one of the team members. Some of the team members were whispering to each other while the team leader was talking, indicated by the observation of *Discussions within small sub-groups*. Many of the participants were engaging in *Closed body language* which is also the most observed behaviour.

In conclusion, the observations and the survey results contradict each other gravely for this team.

---

#### CONCLUSION TRIANGULATION

Summarizing Table 7 and the qualitative observations per team, the results of this study do not support the usage of the observational measures in concert with the survey score. There are discrepancies between the observation results and survey results in most of the teams. This is also visible in the significant positive correlation between the PS ratio and survey

psychological safety which shows that high survey measured psychological safety is related to low levels of observed psychological safety.

However, as the range of survey measured psychological safety is quite small in this study (4.00 – 5.11), some conclusions can be made regarding which behaviours teams with such levels of psychological safety engage in. In three of the four teams the behaviours *Team/TL Providing information* and *TL/TM Active listening* were amongst the top five observed behaviours. This indicates that in most teams the leader was sharing a lot of information, i.e. providing information, and that team members were listening attentively to what the team leader had to say.

## STUDY 2

### RELATIONSHIP OBSERVED PS RATIOS AND TPS

Table 8 shows the relationships between the PS ratios per direction and combined with team psychological safety. Psychological safety's relationship with the PS ratio from all directions combined is significant and the relationship with the PS ratio on TM/TL level is moderately significant. Both relationships are in the expected negative direction. This indicates that a lower PS ratio relates to a higher level of psychological safety, and supports the usage of the observation scheme to measure psychological safety.

	M	SD	TPS
TPS	5.024	.698	
PS ratio (TL/TM)	0.181	.128	.543
PS ratio (TM/TL)	.142	.095	-.657*
PS ratio (TM/TM)	.194	.115	-.179
PS ratio (Team/TL)	.143	.127	-.371
PS ratio (Team/TM)	.236	.097	-.464
PS ratio (all)	.178	.048	-.714**

Table 8: Spearman's Rho correlations TPS with PS ratios  
\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

### RELATIONSHIP ALL OBSERVED BEHAVIOURAL CATEGORIES AND TPS

Table 9 shows the correlations between team psychological safety and the number of behaviours observed in each behavioural category when combining all behavioural directions. *Defensive voice* is significantly correlated with team psychological safety: Teams in which more *Defensive voice behaviours* are observed systemically have lower psychological safety. This is in line with the expected relationship, supporting the use of observing defensive voice behaviours for measuring psychological safety. *Learning and improvement behaviours* have a marginally significant relationship with team psychological safety in the expected direction. In the section below the correlations of specific behaviours and team psychological safety are assessed to find out which behaviours particularly are responsible for these relationships.



	M	SD	TPS
<b>TPS</b>	5.024	.698	
<b>Voice behaviour</b>	22.380	9.820	.393
<b>Defensive voice behaviour</b>	.927	1.062	-.927***
<b>Silence behaviour</b>	7.305	3.226	-.071
<b>Supportive behaviour</b>	48.961	16.351	.214
<b>Unsupportive behaviour</b>	7.445	2.982	-.179
<b>Learning and improvement behaviour</b>	21.226	10.273	.571*
<b>Familiarity behaviour</b>	1.433	2.207	-.433

Table 9: Spearman's Rho correlations TPS with behavioural categories from all levels combined

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP SPECIFIC BEHAVIOURS OBSERVED IN ALL DIRECTIONS COMBINED AND TPS

Table 10 shows the correlations between specific behaviours (combined from all levels) and team psychological safety. Only behaviours that have at least a marginally significant correlation are depicted to make the table more readable. The frequency – the absolute count of how often behaviours occurred during the observational data collection – is also shown. Some behaviours were observed less than 10 times and so the correlations that are derived from this number of observations are not reliable. Only behaviours with at least 10 observations are considered.

	Frequency	M	SD	TPS
<b>TPS</b>		5.024	.698	
<b>Providing information</b>	117	8.661	5.737	.679**
<b>Providing negative feedback (constructively)</b>	14	1.131	1.184	.607*
<b>Correcting others</b>	15	1.411	1.806	.679**
<b>Voicing discontent</b>	12	.648	.818	-.741**
<b>Providing negative feedback (destructively)</b>	4	.150	.395	-.612*
<b>Denying faults or blame others</b>	10	.563	0.553	-.889***
<b>Evading confrontation</b>	3	.177	0.307	-.579*
<b>Showing aggression</b>	1	.037	.0989	-.612*
<b>Facial expression or body language indicates fear</b>	1	.037	.0989	-.612*
<b>Use of inclusive language</b>	8	.299	.791	-.612*
<b>Reacting cold/ignoring a joke</b>	5	.344	.532	-.591*
<b>Verifying progress and performance</b>	75	5.306	2.253	.643*
<b>Accepting feedback</b>	2	.140	.272	-.579*
<b>Asking for ideas, help or solutions</b>	25	1.800	1.379	.577*

Table 10: Spearman's Rho correlations between TPS and specific behaviours from all levels combined

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

*Voicing discontent, Providing information, Correcting others and Providing negative feedback (constructively)* all belong to the category *Voice behaviour*. *Providing information, Correcting others and Providing negative feedback (constructively)* have moderately and marginally significant relationships with team psychological safety in the expected direction. *Voicing discontent* is moderately significantly related to team psychological safety. This relationship is in the unexpected direction. It was expected that when people voice negative opinions, they feel psychologically safe to do so.

*Denying faults or blame others* has a significant negative relationship with team psychological safety. This was expected. This behaviour is part of the *Defensive voice behaviours* which have also been found to be negatively correlated with team psychological safety (see Table 9).

Lastly, *Verifying progress and performance* and *Asking for ideas, help or solutions*, both part of the *Learning and improvement behaviours*, have moderately significant positive relationships with psychological safety. This was expected.

#### RELATIONSHIP OBSERVED TL/TM BEHAVIOURAL CATEGORIES AND TPS

Table 11 shows the relationship between the behavioural categories in a TL/TM direction and team psychological safety. Only for *Unsupportive behaviours* a moderately significant correlation was found. This correlation follows an unexpected direction. Possibly, when looking into the specific behaviours that make up this category, an explanation can be found.

	M	SD	TPS
<b>TPS</b>	5.024	.698	
<b>Voice behaviour</b>	3.059	1.167	-.143
<b>Defensive voice behaviour</b>	.000	.000	
<b>Supportive behaviour</b>	7.737	4.276	.371
<b>Unsupportive behaviour</b>	1.879	1.108	.771**
<b>Learning and improvement behaviour</b>	1.009	1.109	-.493
<b>Familiarity behaviour</b>	.119	.292	-.393

Table 11: Spearman's Rho correlations between TPS and TL/TM behavioural categories

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP SPECIFIC TL/TM BEHAVIOURS OBSERVED AND TPS

Only one moderately significant relationship has been found between *Interrupting* and team psychological safety ( $r = .771$ ;  $p = .072$ ). Two other moderately and marginally significant correlations were found but these behaviours were observed less than 10 times. *Interrupting* is part of the *Unsupportive behaviours*. Also, the direction of *Interrupting* is in the unexpected direction, potentially explaining the unexpected correlation of team psychological safety with *Unsupportive behaviour*.

## RELATIONSHIP OBSERVED TM/TL BEHAVIOURAL CATEGORIES AND TPS

From all behavioural categories, only *Supportive behaviour* has a moderately significant relationship with team psychological safety ( $r = .771$ ,  $p < 0.1$ ). This relationship is in the expected positive direction.

## RELATIONSHIP SPECIFIC TM/TL BEHAVIOURS OBSERVED AND TPS

Only *Agreeing/Responding positively or enthusiastically to input* has been found to have a significant positive relationship with psychological safety ( $r = .829$ ;  $p < 0.05$ ). This correlation is in the expected direction. There were other behaviours that have significant or marginally significant relationships with team psychological safety but these were observed only a few times. Therefore, they are not further considered.

## RELATIONSHIP OBSERVED TM/TM BEHAVIOURAL CATEGORIES AND TPS

Three behavioural categories, when observed in the TM/TM direction, have been found to have significant, moderately significant, and marginally significant relationships with team psychological safety. These are *Unsupportive behaviour*, *Learning and improvement behaviour*, and *Familiarity behaviour*, respectively (see Table 12). While the relationship with *Unsupportive behaviour* is in the expected direction, this is not the case for *Learning and improvement behaviour* and *Familiarity behaviour*.

	M	SD	TPS
TPS	5.024	.698	
Voice behaviour	5.428	5.029	-.393
Defensive voice behaviour	.205	.541	-.204
Supportive behaviour	22.101	10.059	-.071
Unsupportive behaviour	4.517	2.718	-.821***
Learning and improvement behaviour	2.497	2.790	-.714**
Familiarity behaviour	.075	.198	-.612*

Table 12: Spearman's Rho correlations TPS and TM/TM behavioural categories

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

## RELATIONSHIP SPECIFIC TM/TM BEHAVIOURS OBSERVED AND TPS

Three specific behaviours in the TM/TM direction have been found to be related to team psychological safety (see Table 13). *Interrupting* has a significant negative relationship with team psychological safety, explaining the negative relationship of *Unsupportive behaviour* with team psychological safety. This direction of the relationship was expected. *Informing about issues or mistakes* and *Speaking up with ideas* also have marginally significant and significant negative correlations with team psychological safety. This is striking as these behaviours should have indicated higher psychological safety based on theory.

	Frequency	M	SD	TPS
<b>TPS</b>		5.024	.698	
<b>Providing information</b>	7	.527	.854	-.749**
<b>Providing positive feedback</b>	4	.279	.544	-.579*
<b>Providing help or solutions</b>	5	.382	.803	-.579*
<b>Voicing discontent</b>	3	.242	.534	-.579*
<b>Interrupting</b>	65	3.696	2.455	-.857***
<b>Reacting cold/ignoring a joke</b>	2	.075	.198	-.612*
<b>Accepting feedback</b>	1	.037	.099	-.612*
<b>Asking for ideas, help or solutions</b>	4	.214	.367	-.668*
<b>Informing about issues or mistakes</b>	10	.815	1.304	-.593*
<b>Speaking up with ideas</b>	10	.397	.877	-.802***
<b>Making or laughing about a joke</b>	2	.075	.198	-.612*

Table 13: Spearman's Rho correlations TPS and TM/TM specific behaviours

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP OBSERVED TEAM/TL BEHAVIOURAL CATEGORIES AND TPS

Looking at the Team/TL direction, three behavioural categories are related to team psychological safety: *Voice behaviour*, *Defensive voice behaviour* and *Learning and improvement behaviour* (see Table 14). All of these relationships follow the expected direction.

	M	SD	TPS
<b>TPS</b>	5.024	.698	
<b>Voice behaviour</b>	4.949	2.174	.657*
<b>Defensive voice behaviour</b>	.143	.348	-.655*
<b>Silence behaviour</b>	1.812	1.237	.143
<b>Supportive behaviour</b>	3.350	1.718	.143
<b>Unsupportive behaviour</b>	.000	.000	
<b>Learning and improvement behaviour</b>	7.048	5.861	.771**
<b>Familiarity behaviour</b>	.194	.313	-.439

Table 14: Spearman's Rho correlations TPS and Team/TL behavioural categories

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP SPECIFIC TEAM/TL BEHAVIOURS OBSERVED AND TPS

Table 15 shows the relationship between team psychological safety and various specific behaviours in the Team/TL direction: *Providing information*, *Verifying progress and performance*, and *Asking for ideas, help or solutions*. All of them are positive, which was expected.

	Frequency	M	SD	TPS
<b>TPS</b>		5.024	.698	
<b>Providing information</b>	25	2.345	1.722	.886***
<b>Denying faults or blame others</b>	2	.142	.348	-.655*
<b>Facial expression or body language indicates disengagement</b>	4	.284	.696	-.655*
<b>Verifying progress and performance</b>	32	2.948	2.251	.771**
<b>Asking for ideas, help or solutions</b>	9	.926	.817	.841***
<b>Speaking up with ideas</b>	13	1.451	1.850	.754**

Table 15: Spearman's Rho correlations TPS and Team/TL specific behaviours

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP OBSERVED TEAM/TM BEHAVIOURAL CATEGORIES AND TPS

Two of the behavioural categories in the Team/TM direction have a significant and marginally significant relationship with team psychological safety: *Voice* and *Defensive voice behaviour*. Both are in the expected direction (see Table 16).

	M	SD	TPS
<b>TPS</b>	5.024	.698	
<b>Voice behaviour</b>	8.196	5.295	.607*
<b>Defensive voice behaviour</b>	.472	1.073	-.802***
<b>Silence behaviour</b>	5.751	3.172	.071
<b>Supportive behaviour</b>	9.169	6.110	-.321
<b>Unsupportive behaviour</b>	.000	.000	
<b>Learning and improvement behaviour</b>	10.227	6.999	.357
<b>Familiarity behaviour</b>	.988	1.739	-.433

Table 16: Spearman's Rho correlations TPS and Team/TM behavioural categories

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

#### RELATIONSHIP SPECIFIC TEAM/TM BEHAVIOURS OBSERVED AND TPS

Not one of the specific behaviours in the Team/TM direction is related to team psychological safety and observed more than 10 times. This means that the results from Table 16 might also be questionable.

#### CONCLUSION OBSERVED BEHAVIOURS AND TPS

In conclusion, when looking at the relationship between the observed behaviours when summarized in a PS ratio with team psychological safety, moderate significance has been found. This provides some support for measuring psychological safety using observations next to surveys.

Table 17 summarizes the correlations found between specific observed behaviours and team psychological safety. Overall, several significant relationships have been found. This indicates that using observations can be useful as an extension of researching psychological safety with survey-based methods. The observations uncover what kind of behaviours happen more in teams with higher and teams with lower psychological safety.

Significance	Positive relationships	Negative relationships
p < 0.05	TM/TL Agreeing	Denying fault or blame others (all)
	Team/TL Providing information	TM/TM Interrupting
	Team/TL Asking for ideas, help or solutions	TM/TM Speaking up with ideas
p < 0.1	Providing information (all)	Voicing discontent (all)
	TL/TM Interrupting	
	Correcting others (all)	
p < 0.2	Team/TL Verifying progress and performance	
	Asking for ideas, help or solutions (all)	TM/TM Informing about issues or mistakes
	Providing negative feedback (constructively) (all)	
	Verifying progress and performance (all)	

Table 17: Specific behaviours associated with TPS categorized by direction of the association

#### TEAM PERFORMANCE AND TEAM PSYCHOLOGICAL SAFETY

No significant relationship is found between team performance and team psychological safety ( $r = -.600$ ;  $p = .208$ ). However, the correlation does approach marginal significance.

#### TRIANGULATING THE TEAM-SPECIFIC RESULTS

This section looks at the quantitative key findings for each team separately and compares them with qualitative observations. Table 18 summarizes the key quantitative findings.

#### TEAM 1

Team 1 has moderate survey-measured team psychological safety. This is also reflected in the moderate PS ratio, and it can be seen in the top 5 most observed behaviours. Only one of these behaviours, is a behaviour that is related to low psychological safety: *TM/TM Interrupting*

	TPS	PS ratio	5 most observed behaviours
Team 1	4.39	0.2	<ol style="list-style-type: none"> <li>1. TM/TM Agreeing</li> <li>2. TM/TM Active listening</li> <li>3. TM/TM Interrupting</li> <li>4. TM/TL Active listening</li> <li>5. TL/TM Agreeing</li> </ol>

<b>Team 2</b>	5.25	0.24	<ol style="list-style-type: none"> <li>1. TM/TM Active listening</li> <li>2. Team/TM Making or laughing about a joke</li> <li>3. TM/TL Active listening</li> <li>4. Team/TL Verifying progress and performance</li> <li>5. TM/TM Agreeing</li> </ol>
<b>Team 3</b>	4.72	0.16	<ol style="list-style-type: none"> <li>1. TM/TM Active listening</li> <li>2. TM/TM Agreeing</li> <li>3. Team/TM Sharing future plans</li> <li>4. TL/TM Agreeing</li> <li>5. Team/TM Closed body language</li> <li>5. TM/TM Asking for further clarification</li> </ol>
<b>Team 4</b>	3.94	0.23	<ol style="list-style-type: none"> <li>1. TM/TM Agreeing</li> <li>2. TM/TM Active listening</li> <li>3. Team/TM Closed body language</li> <li>3. Team/TM Sharing future plans</li> <li>5. Team/TM Informing about issues as mistakes</li> </ol>
<b>Team 5</b>	5.61	0.15	<ol style="list-style-type: none"> <li>1. TM/TM Agreeing</li> <li>1. Team/TM Providing information</li> <li>1. Team/TM Informing about issues and mistakes</li> <li>4. TM/TL Agreeing</li> <li>5. TM/TL Active listening</li> </ol>
<b>Team 6</b>	5.38	0.17	<ol style="list-style-type: none"> <li>1. Team/TM Providing information</li> <li>2. TM/TM Active listening</li> <li>3. Team/TM Informing about issues or mistakes</li> <li>4. TL/TM Active listening</li> <li>5. Team/TL Verifying progress and performance</li> </ol>
<b>Team 7</b>	5.88	0.1	<ol style="list-style-type: none"> <li>1. TM/TM Agreeing</li> <li>2. TM/TM Active listening</li> <li>2. TM/TL Agreeing</li> <li>4. Team/TM Speaking up with ideas</li> <li>5. TL/TM Agreeing</li> </ol>

*Table 18: Overview key figures for each team*

During the meeting, there was a lot of talk about the progress of the team and members freely gave their opinions or talked to the team about issues they encountered. This indicated high psychological safety. What stood out in this team was that the team talked in two languages simultaneously: German and Dutch. While they decided to speak Dutch at the beginning of the meeting, some of the members quickly switched to German again. In the end, the language that was spoken switched constantly. Potentially, people that speak better German do not feel safe when the meeting is conducted in Dutch and vice versa. Also, when two people start to speak German, it can come across to the rest of the team as if they should

not be included in that conversation. This type of behaviour could explain why the level of psychological safety is not higher.

To conclude, the survey, quantitative and qualitative observations all point towards Team 1 having moderate psychological safety. Moreover, the quantitative and qualitative observations shed light on why only a moderate level of psychological safety was measured with the survey. This supports the usage of the three methods collectively.

---

#### TEAM 2

In Team 2, a high level of team psychological safety was measured with the survey. This does not match the PS ratio of 0.24, which is the highest of this study and would indicate that the team has the lowest psychological safety. However, the top five observed behaviours do not show any behaviours that are theorized to be negatively related to psychological safety.

Looking at this team qualitatively, the meeting consisted mainly of the team leader communicating towards the team members and most of the team members did not interact with the team leader or other team members. However, they were listening so this passive behaviour does not necessarily mean that there was low psychological safety. Moreover, from a methodological perspective this team was hard to observe because there were a lot of team members (14). This leads to the observer easily overlooking behaviour of some participants. Both of these factors could have distorted the PS ratio.

In conclusion, the PS ratio contradicts the surveyed team psychological safety score, while the top five behaviours support the surveyed score. Qualitatively, it was hard to observe whether the team was psychologically safe or not as team members were quiet for most of the (relatively short) meeting. All things considered, following the triangulation of this team, it remains questionable whether it is possible to use the three methodologies together.

---

#### TEAM 3

Team 3 has moderately high survey-measured team psychological safety, which is also reflected in the low PS ratio. Only one of the top five behaviours is one that has been theorized to be indicative of low psychological safety: *Closed body language*. However, during the statistical analysis for this behaviour no significant relationship with psychological safety has been found, so potentially this behaviour does not influence psychological safety.

From a qualitative perspective, this team had a very methodical meeting style where they started with discussing the progress each team member had made and ended with each team member's plan for the day. This is also reflected in the amount of times *Sharing future plans* was observed. It was also acceptable for team members to admit to not having completed their tasks yet, and the team tried to search for solutions together. Such behaviour would be indicative of psychological safety.



Concluding, the three methods – the survey, the quantitative and the qualitative observations – agree on the moderately high level of psychological safety, supporting the use of the three methodologies in concert.

---

#### TEAM 4

Team 4 is the only team in this study that has a psychological safety score of under 4, indicating low psychological safety. On the other hand, the team has the second to lowest PS ratio and the top five observed behaviours show only one behaviour indicative of low psychological safety, namely *Closed body language*. The qualitative observation might explain this discrepancy.

For quite a large part of the meeting the team members were discussing about a new colleague who seemed to be part of the team officially but not really part of the group. Team members were saying that they do not feel it is necessary to include her more, that it is hard to talk to her because she only works part-time and also that they do not feel they have something in common so they do not feel the need to talk to the person on a personal level. The team member that they were talking about was also not at the recorded meeting and they were talking about future activities where they considered not to invite her either. This kind of behaviour, talking about someone behind their back in an unpleasant way, could implicate low team psychological safety. People that are present in the meeting might start wondering if the team members are talking about them as well when they are not there and, therefore, might not feel safe to speak their minds completely.

All in all, while the survey and qualitative observations point towards low psychological safety, the quantitative observations show high psychological safety. This indicates that the observation scheme might not yet be fully suitable for use in psychological safety research.

---

#### TEAM 5

Team 5 has high psychological safety, both according to the survey and according to the PS ratio. The top 5 observed behaviours also encompass only behaviours indicative of high psychological safety.

From a qualitative perspective, the meeting went as follows: The leader would steer the conversation, determining what topics are discussed based on a pre-made agenda. Then he would invite all team members to comment on the topics, experiences they have had, issues they encountered and potential solutions for these issues. This meeting style seems to align with a psychologically safe environment. No striking or outstanding situations were encountered.

To conclude, the three methods agree on a high level of psychological safety for this team.

---

#### TEAM 6

This team has a relatively high team psychological safety score of 5.38, which is also reflected in the moderately low PS ratio. The top five observed behaviours include only behaviours theorized to be indicative of high psychological safety. One of the five behaviours – Team/TL Verifying progress and performance – has also statistically been found to be positively related to psychological safety.

Qualitatively, the meeting had a friendly ambiance and jokes were made between team members themselves and the team leader and team members. One of the team members was rather quiet throughout the meeting. The team was focused on communicating about issues, which is reflected in the behaviour Informing about issues and mistakes, and trying to find solutions for them together.

All in all, the three methods of observation seem to align. The qualitative observation, however, revealed no exceptional results.

---

#### TEAM 7

Team 7 has the highest survey measured psychological safety and the lowest PS ratio of this study. Also, the top five behaviours show only positive behaviours.

The top five behaviours align with the qualitative observations: During the meeting, most of the time one person would suggest a topic to talk about or a solution for a problem – *Speaking up with ideas* – and the rest would listen to this person and, most of the time, agree with the idea. This seems to constitute a psychologically safe environment.

In conclusion, all methods agree on the high psychological safety in this team.

---

#### CONCLUSION TRIANGULATION

Considering all teams together, the triangulation of the findings, overall, shows some alignment between the three methods of analysis. Also, there were several behaviours that were observed in multiple teams, these being *Active listening*, *Agreeing* and *Informing about issues or mistakes*. However, these behaviours were observed irrespective of the surveyed psychological safety score of the team, indicating that they occur in teams with different levels of psychological safety.

Only two of the behaviours that have been found to be significantly related to psychological safety in the quantitative analyses are included in the top five most observed behaviours: *Team/TL Verifying progress and performance* and *TM/TM Interrupting*. *Team/TL Verifying progress and performance* has been one of the top five behaviours in the team with the highest surveyed psychological safety, while *TM/TM Interrupting* has been one of the top five behaviours in the team with the second to lowest surveyed psychological safety.

All in all, the triangulation supports the use of the three methods of this research in concert. However, it is also indicated that the observation scheme could be improved to some degree to include more relevant behaviour, e.g. gossiping.

## STUDY 3

The analysis of study 3 starts with a quantitative analysis of the relationship between the observed behaviours and psychological safety, and squad performance and psychological safety.

## RELATIONSHIP INDIVIDUAL PSYCHOLOGICAL SAFETY AND SQUAD PSYCHOLOGICAL SAFETY

Since the SPS data are from the second meeting and the observational and SP data are from the third meeting, it needs to be checked whether the SPS data from the second meeting are still relevant at the third meeting. For this, scores of IPS from meeting 2 and meeting 3 are compared. There is a significant relationship found between these two scores ( $r = .48$   $p = .005$ ) indicating that individual psychological safety levels have not changed between the two meetings. It is assumed that this translates to the team-level, so the SPS scores from the second meeting can be compared to the rest of the data from the third meeting.

## RELATIONSHIP OBSERVED PS RATIOS AND SURVEYED SPS

Firstly, the PS ratio based on the observed behaviour is compared with SPS.

No significant relationships have been found. However, the association between the Team/TM PS ratio and surveyed SPS approaches marginal significance ( $r = 0.600$ ;  $p = 0.208$ ). However, the direction of this correlation indicates that the higher the Team/TM PS ratio the higher the level of SPS is which is striking. A lower PS ratio should indicate higher SPS based on the theoretical distribution of behaviours into behaviours affecting PS positively and negatively.

## RELATIONSHIP ALL OBSERVED BEHAVIOURAL CATEGORIES AND SPS

Next the behavioural instances counted at both levels combined in the behaviour categories are compared to SPS. Table 19 shows the Spearman-Rho correlations for all behaviour categories with SPS. Two of the behavioural categories are marginally significant, namely *Voice Behaviour* and *Unsupportive behaviour*. Both seem to have a positive relationship with SPS, indicating that squads that have a lot of *Voice Behaviour* and *Unsupportive Behaviour* also have higher SPS. However, theoretically, *Unsupportive behaviour* should be negatively associated with SPS. Below, the distinct behaviours that make up each category are analysed.

## RELATIONSHIP SPECIFIC BEHAVIOURS OBSERVED IN BOTH DIRECTIONS COMBINED AND SPS

Table 20 summarizes the distinct behaviours that had at least a marginally significant relationship with the team-level psychological safety score and shows the total number of times the behaviour has been observed in this study.

Because *Asking for feedback* has only been observed 3 times, making the chance that the effect is just coincidental very high, this behaviour is not further considered.

	M	SD	SPS
SPS	5.715	0.497	
Voice behaviour	16.651	2.910	.657*
Defensive voice behaviour	2.620	2.626	.314
Silence behaviour	2.191	1.461	.200
Supportive behaviour	24.608	4.776	.029
Unsupportive behaviour	6.955	4.251	.657*
Learning and Improvement behaviour	18.351	4.259	-.257
Familiarity behaviour	4.185	6.180	0.429

Table 19: Spearman's Rho correlations SPS and behavioural categories (TM/TM and Team/TM summed)  
\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

	Frequency	M	SD	SPS
SPS		5.715	0.497	
Providing negative feedback (constructively)	72	1.705	0.955	-.771**
Sharing future plans	73	1.695	1.018	.886***
Discussions within small subgroups	12	0.363	0.508	.941***
Reacting cold/ignoring a joke (only TM/TM)	32	1.061	1.985	.771**
Asking for feedback	3	0.055	0.064	-.638*

Table 20: Spearman's Rho correlations SPS and specific behaviours that show some significance (TM/TM and Team/TM summed)  
\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

While *Sharing future plans*, *Discussions with small sub-groups* and *Reacting cold/ignoring a joke* seem to associate with SPS positively, *Providing negative feedback (constructively)* seems to associate with SPS negatively. The direction of three of these associations are striking.

Both, *Discussions with small sub-groups* and *Reacting cold/ignoring a joke* fall under the category of *Unsupportive behaviours* and, therefore, are probably the reason for the significant relationship found between this behavioural category and SPS.

Lastly, *Providing negative feedback (constructively)* was expected to be positively related to psychological safety. However, the Spearman-Rho correlation indicates otherwise. This behaviour falls under the category of *Voice behaviours* and could, thereupon, explain the significant relationship of this behavioural category with SPS.

#### RELATIONSHIP OBSERVED TM/TM BEHAVIOURAL CATEGORIES AND SPS

This section presents the Spearman-Rho correlations between the behaviours that have been observed between individual squad members and SPS.

From all behaviours only *Familiarity behaviour* is marginally significant ( $r = .657$ ,  $p = .156$ ). This indicates that there is a positive relationship between the familiarity behaviour individual squad members engage in with each other and the level of squad psychological safety they feel. Subsequently, the behaviours that make up each category are analysed.

## RELATIONSHIP SPECIFIC TM/TM BEHAVIOURS OBSERVED AND SPS

Table 21 shows the behaviours on a TM/TM level that have a significant relationship with survey-measured SPS. Behaviours that have only been observed a handful of times, however, will not be considered further as the measured effect is probably not reliable.

It has been found that, for interactions between individual squad members, the behaviours *Asking for further clarification*, *Reacting cold/ignoring a joke* and *Making or laughing about a joke* are positively associated with psychological safety, and that the behaviours *Sharing procedures, knowledge and experience* and *Verifying progress and performance* are negatively associated with psychological safety.

This is contradicting the expectations and the theoretical categorization of the behaviours. It was expected that *Reacting cold/ignoring a joke* would be negatively associated with squad psychological safety, and that *Sharing procedures, knowledge and experience* and *Verifying progress and performance* would be positively associated with squad psychological safety.

The behaviours *Asking for further clarification* and *Making or laughing about a joke*, which are positively associated with psychological safety, indicate that squads in which squad members engage in these behaviours towards other squad members have higher levels of squad psychological safety.

Interestingly, many of the specific behaviours identified to be associated with SPS fall under behavioural categories that have not been rendered related to SPS. From the category *Familiarity behaviour*, that does have a relationship with SPS, only the specific behaviour *Making or laughing about a joke* is related to SPS.

	Frequency	M	SD	SPS
SPS		5.715	0.497	
Asking for further clarification	143	3.576	1.144	.714*
Asking a question (other)	2	0.049	0.120	-.655*
Voicing discontent	1	0.025	0.060	-.655*
Showing aggression	2	0.050	0.120	-.655
Sharing procedures knowledge and experience	24	0.614	0.759	-.829***
Acknowledging achievements/congratulating	4	0.080	0.115	-.638*
Delegating tasks	3	0.134	0.328	.655*
Use of inclusive language	3	0.074	0.180	.655*
Discussions within small subgroups (only TM/TM)	4	0.115	0.215	.676*
Reacting cold/ignoring a joke (only TM/TM)	32	1.060	1.985	.771**
Verifying progress and performance	31	0.671	0.509	-.943***
Making or laughing about a joke	71	1.671	1.546	.657*

Table 21: Spearman's Rho correlations SPS and specific behaviours that show some significance (only TM/TM)  
\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

## RELATIONSHIP OBSERVED TEAM/TM BEHAVIOURAL CATEGORIES AND SPS

Table 22 shows the Spearman-Rho correlations for the behavioural categories on the Team/TM level with SPS.

As can be seen below, on the Team/TM level one behavioural category is significantly positively related to SPS, being *Unsupportive behaviour*. Another behavioural category approaching marginal significance is *Defensive Voice behaviour*, also relating positively to SPS. Both of these are striking as they are in an unexpected direction. Again, below the specific behaviours of each category are analysed.

	M	SD	SPS
SPS	5.715	0.497	
Voice behaviour	7.781	2.128	.257
Defensive voice behaviour	1.852	2.144	.600
Silence behaviour	2.191	1.461	.200
Supportive behaviour	6.659	2.188	-.029
Unsupportive behaviour	0.516	0.884	.845***
Learning and improvement behaviour	9.176	2.130	-.257
Familiarity behaviour	2.514	4.818	.257

Table 22: Spearman's Rho correlations SPS and behavioural categories (only Team/TM)  
\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

## RELATIONSHIP SPECIFIC TEAM/TM BEHAVIOURS AND SPS

Table 23 summarizes the distinct behaviours on a Team/TM level that have a significant relationship with the team-level psychological safety score and shows the total number of times the behaviour has been observed in this study. However, for behaviours that have been observed less than 10 times there is a high probability that the effect measured was just by chance. Therefore, these are not further considered.

On the level of squad member towards the whole squad, only the behaviours *Denying faults or blame others* and *Sharing future plans* has been found to be associated with squad psychological safety and in a positive direction. This is striking for the behaviour *Denying faults or blame others* as it should be a behaviour that is negatively associated with psychological safety. On the other hand, the positive effect of *Sharing future plans* was expected and can indicate that, based on this sample, squads in which squad members share their future plans with the whole squad have higher psychological safety.

Looking back at the behavioural categories on the Team/TM level that are associated with SPS, only for *Defensive Voice behaviour* a specific behaviour has been identified to be associated with SPS: *Denying faults or blame others*. Regarding the category *Unsupportive behaviour*, it can be seen in the table below that two behaviours have a significant relationship but with very little instance observed and, therefore, they are not further discussed here. This could indicate that also the relationship between the behavioural category *Unsupportive behaviour* and SPS is questionable.

	Frequency	M	SD	SPS
<b>SPS</b>		5.715	0.497	
<b>Asking a question (other)</b>	1	0.045	0.109	.655*
<b>Denying fault or blame others</b>	42	1.187	1.249	.657*
<b>Evading confrontation</b>	3	0.089	0.110	.820***
<b>Sharing future plans</b>	67	1.508	0.932	.886***
<b>Interrupting</b>	6	0.268	0.656	.655*
<b>Discussion within small sub-groups</b>	8	0.248	0.406	.778**
<b>Asking for feedback</b>	3	0.055	0.064	-.638*

Table 23: Spearman's Rho correlations SPS and specific behaviours that show some significance (only Team/TM)

\*  $p < 0.2$ , \*\*  $p < 0.1$ , \*\*\*  $p < 0.05$

### CONCLUSION OBSERVED BEHAVIOURS AND SPS

Concluding the multi-level analyses above, overall, the relationship between SPS and the ratio of behaviours being positive and negative for psychological safety seems weak. An indication of a relationship exists for the PS ratio based on behaviour that squad members exert to the squad as a whole, however, this relationship follows an unexpected direction. Nevertheless, various observable behaviours were found that might relate to psychological safety. Table 24 below summarizes these behaviours categorized by the direction of their relationship with SPS and their significance level.

	Positive relationship	Negative relationship
p < 0.05	Team/TM Sharing future plans	TM/TM Sharing procedures, knowledge and expertise
	Sharing future plans (all)	TM/TM Verifying progress and performance
	Discussions within small sub-groups (all)	
p < 0.1	TM/TM Making or laughing about a joke	Providing negative feedback (constructively) (all)
	TM/TM Reacting cold/ignoring a joke	
	Reacting cold/ignoring a joke (all)	
	TM/TM Asking for further clarification	
p < 0.2	Team/TM Denying faults or blame others	

Table 24: Specific behaviours associated with SPS categorized by direction of the association

### SQUAD PERFORMANCE AND SQUAD PSYCHOLOGICAL SAFETY

Squad performance has been assessed by the squad members themselves as well as a number of "experts". Spearman Rho correlation analyses showed that expert-rated squad performance was not related to self-rated squad performance ( $r = .29$ ;  $p = .58$ ).

It is assumed that the score given by the experts is more reliable as it is more objective than the score the squad gave itself. So, further analysis considers only the expert opinion on SP.

Correlating expert-rated squad performance with SPS results in a marginally significant positive relationship ( $r = .609$   $p = 0.2$ ). Squads that have higher psychological safety, also have higher squad performance. This means that squad psychological safety could be relevant to squad performance.

#### TRIANGULATING THE SQUAD-SPECIFIC RESULTS

In this section for each squad the outcomes of the survey, the quantitative observation and the qualitative observation are compared. Table 25 gives an overview of the key findings for each squad.

	Survey SPS	Observed SPS ratio	5 most observed behaviours
<b>Squad 1</b>	5.92	0.14	<ol style="list-style-type: none"> <li>1. TM/TM Active listening</li> <li>2. TM/TM Agreeing</li> <li>3. TM/TM Interrupting</li> <li>4. Team/TM Providing information</li> <li>5. Team/TM Sharing procedures, knowledge and experience</li> </ol>
<b>Squad 2</b>	5.82	0.09	<ol style="list-style-type: none"> <li>1. TM/TM Active listening</li> <li>2. TM/TM Agreeing</li> <li>3. Team/TM Informing about issues or mistakes</li> <li>4. Team/TM Sharing procedures, knowledge and experience</li> <li>5. TM/TM Asking for further clarification</li> </ol>
<b>Squad 3</b>	6.19	0.18	<ol style="list-style-type: none"> <li>1. Team/TM Making or laughing about a joke</li> <li>2. TM/TM Agreeing</li> <li>3. TM/TM Reacting cold/ignoring a joke</li> <li>4. TM/TM Asking for further clarification</li> <li>4. TM/TM Making or laughing about a joke</li> <li>4. Team/TM Speaking up with ideas</li> </ol>
<b>Squad 4</b>	5.38	0.15	<ol style="list-style-type: none"> <li>1. TM/TM Active listening</li> <li>1. TM/TM Agreeing</li> <li>3. TM/TM Interrupting</li> <li>4. Team/TM Informing about issues or mistakes</li> <li>5. Team/TM Providing information</li> </ol>
<b>Squad 5</b>	4.88	0.18	<ol style="list-style-type: none"> <li>1. TM/TM Active listening</li> <li>2. TM/TM Agreeing</li> <li>3. TM/TM Interrupting</li> <li>4. Team/TM Speaking up with ideas</li> <li>5. Team/TM Sharing procedures, knowledge and experience</li> </ol>



<b>Squad 6</b>	6.10	0.33	<ol style="list-style-type: none"> <li>1. TM/TM Interrupting</li> <li>2. TM/TM Agreeing</li> <li>3. TM/TM Active listening</li> <li>4. TM/TM Disagreeing</li> <li>5. Team/TM Sharing procedures, knowledge and experience</li> </ol>
----------------	------	------	--

*Table 25: Overview key figures for each squad*

#### SQUAD 1

This squad has a relatively high SPS level of 5.92. This is also reflected in the low SPS ratio that came from the observations, which shows that the squad engages in far more behaviours that indicate high psychological safety than behaviours that indicate low psychological safety. This can also be seen in the top five behaviours that were observed in the squad: four of them may indicate a high level of psychological safety.

However, also one behaviour that should have negative repercussions for psychological safety has been observed very often, being squad members interrupting other squad members.

Based on our qualitative observations, the observed retrospective meeting was very thoroughly structured to discuss team-level as well as individual-level achievements and each member has been asked their personal opinion on several instances. Moreover, there was a strong emphasis on finding agreement on the subjects discussed. However, the opinion of some members seemed to overshadow the opinions of others which was also reflected in some members being a bit hesitant when asked their opinion. This could explain why the level of psychological safety was not higher.

In conclusion, squad 1 shows high psychological safety using all three methods of the analysis: the survey score, the observed PS ratio, and the qualitative observations. This points to the strength of the combined usage of these three methods.

#### SQUAD 2

The second squad, also, has a relatively high survey-measured SPS. Additionally, their observed SPS ratio indicates a high level of psychological safety. This again indicates that the observed SPS ratio and the surveyed SPS seem to align in this study. The top five behaviours in this squad are all behaviours indicative of high psychological safety.

Zooming in on what happened during the observed meeting, it seemed a quite obligatory meeting and people did not seem particularly keen to participate in it. There was one person who chaired the meeting and often members were talking directly to him or to another person who seemed to have a supporting role as well. There was little interaction between other squad members. Nevertheless, squad members were listening and hardly showed disengagement behaviours.

In conclusion, for squad 2, while the intensity of interaction between members was limited, no psychological safety depleting behaviours were observed either. So, the qualitative observations somewhat align with the quantitative results. It is interesting to see that the passivity encountered in this squad does not necessarily mean that the squad has low psychological safety.

---

#### SQUAD 3

This squad has the highest surveyed SPS-score of all squads in this study. The observed SPS ratio is also quite low. In this squad three behaviours were observed exactly the same amount of times, so all three are included in the top five, resulting in a top six. It is notable that three of the behaviours are concerning joke-making or laughing about jokes. This was also seen during the qualitative observations. The squad was very much entertaining each other and making jokes and when somebody made a joke, the next person would make another joke as a response, resulting in a very relaxed and friendly meeting ambiance. Even jokes at the expense of other squad members were common and mostly appreciated. Indeed, it seemed that the members of this squad were acting more like friends than ‘just’ like colleagues. Apparently, this positive climate translates to higher psychological safety as can be seen from the surveyed SPS score. Interestingly, this is the only squad that does not have *Active listening* in the top five of their behaviours.

In conclusion, the results of the three methods corroborate the findings that squad 3 is a highly psychologically safe squad. The positive effect of the relaxed ambiance with its many jokes is especially mentionable.

---

#### SQUAD 4

This squad has a moderately high surveyed SPS score. This is also reflected in the relatively low observed SPS ratio. Regarding both measures this study stands in the middle of all other squads in this study, indicating some alignment between observed SPS and surveyed SPS. Looking at the top five most observed behaviours, only with this squad *Providing information* towards the squad as a whole is included.

This aligns with the qualitative observations of this squad: their meeting seemed very formal. People mainly elaborated on their progress, concerns and ideas to mitigate problems. Thus, a lot of *Providing information* occurred. It is striking that *Interrupting* has been coded so often, as from a qualitative perspective it did not seem like there were elaborate discussions on topics but there was more of a planned turn-taking as to who was speaking. Possibly, the high rate of *Interrupting* is due to one person who often added his thoughts while others were talking, as well as the chairperson who interrupted in order to move the discussion to the next agenda point. Overall, the squad seemed to focus on getting everything done quickly.

In conclusion, the observed SPS ratio and the surveyed PS ratio seem to align. The qualitative analysis also aligns with this to the extent that it also showed a moderately high level of

psychological safety. However, on some aspects the qualitative analysis deviates from the top five observed behaviours.

---

#### SQUAD 5

Squad 5 has the lowest surveyed SPS score of all squads in this study. Still, the score is on the positive end of the scale, indicating that some psychological safety persists. The observed SPS ratio indicates moderate levels of psychological safety. This can also be seen from the five most observed behaviours, where only *Interrupting* is included as a potentially negative behaviour for psychological safety.

During the meeting, two squad members had hefty discussions about the work of one of the two. One squad member was arguing that the other repeatedly did not do his work correctly, while the other was blaming the first squad member for not giving him sufficient information to execute the tasks properly. This argument resulted in tension between the two members but also the climate of the whole meeting got aggravated. Moreover, this squad was one of the few in which the behaviours *Denying faults or blame other* and *Showing aggression* were scored repeatedly, albeit not enough to come up in the top five behaviours. So, to some extent the quantitative observation shows what has been observed qualitatively as well.

In conclusion, for this squad, initially, some differences between the survey-based results and the observation-based quantitative results can be found. However, when considering the qualitative observations in concert with the quantitative observations for behaviours other than the top five, some alignment can be found. All in all, while the qualitative and survey-based results align, the quantitative observations, particularly the PS ratio, are misleading for this squad.

---

#### SQUAD 6

Squad 6 has the highest psychological safety in this study based on the survey but the lowest psychological safety based on the observations. *Interrupting*, a negative behaviour, has been scored the most in the quantitative observation which explains why the observed SPS ratio is so high.

Regarding the qualitative analysis, one squad member was very outspoken while the rest was relatively quiet. This one member had very strong opinions that were often negative and he defended his viewpoints aggressively. He often interrupted other people and talked over them to emphasize his point, a form of intrusive interruption. The fact that this behaviour does not negatively affect psychological safety, when looking at the survey-based score, is striking.

In conclusion, for this squad the surveyed results contradict with the results from the observational analysis.

---

## CONCLUSION TRIANGULATION

Looking back at the triangulation of results for all squads, some alignment between the three methods can be found. Sometimes, both quantitative results aligned but not the qualitative analysis and sometimes only the observational results aligned but not the survey results. Only for two of the six squads the results of the three methods fully aligned with each other.

However, the differences between the squads in terms of psychological safety are minimal: all teams have a relatively high score for psychological safety and all squads, with the exception of the last squad, have a relatively low PS ratio. When excluding the last squad, there seems to be major alignment between the survey results, PS ratio and the most observed behaviours: Each team engages in *TM/TM Agreeing*, four of the five team engage in *TM/TM Active listening* and three of the five engage in *TM/TM Interrupting*. Moreover, all teams engage in some form of information sharing through behaviours such as *Informing about issues and mistakes*, *Sharing procedures, knowledge and experience*, *Providing information* or *Speaking up with ideas*. These behaviours, thus, seem to be related to teams that have relatively high psychological safety. From the qualitative observations, it can be concluded that there can be different styles or ambiances in meetings, which all are related to relatively high psychological safety. Some squads had rather rigid, practical and formal meetings while other squads were more relaxed, social and frenzied.

Comparing to the statistical analyses, only two of the behaviours that were found to relate to psychological safety are actually included in the top five behaviours: *TM/TM Making or laughing about a joke* and *TM/TM Asking for further clarification*. Both of these have a positive relationship with psychological safety and have been observed in the squad with the highest surveyed psychological safety.

## STUDY 4

In this study, squad 3 from study 3 is re-observed with the help of the computer programme The Observer XT. The findings are split into conceptual and methodological findings.

---

## CONCEPTUAL FINDINGS

Table 26 compares the SPS ratio and 5 most observed behaviours based on the naked-eye coding and the coding in The Observer XT, and shows the five longest lasting behaviours. When looking at the 5 most observed behaviours, there is a lot of overlap. However, coding in The Observer has revealed that Squad 3 also used a lot of inclusive language during the meeting and that team members were listening actively to their colleagues.

Considering the five longest lasting behaviours new insights can be found. Apparently, squad members engaged in *Closed body language* for a long time during the meeting. Recalling the high team psychological safety score, this behaviour does not necessarily seem to relate to psychological safety. Furthermore, the squad spent much time on *Speaking up with ideas* and *Informing about issues or mistakes* towards the whole team. This fits the purpose of the

meeting, as it was a retrospective. More importantly, it fits the high level of psychological safety since both of these behaviours can be termed interpersonally risky.

	Naked-eye	The Observer XT
<b>SPS ratio</b>	0.18	0.19
<b>5 most observed behaviours</b>	1. Team/TM Making or laughing about a joke	1. TM/TM Active listening
	2. TM/TM Agreeing	2. TM/TM Agreeing
	3. TM/TM Reacting cold/ignoring a joke	3. TM/TM Reacting cold/ignoring a joke
	4. TM/TM Asking for further clarification	5. Team/TM Use of inclusive language
	4. TM/TM Making or laughing about a joke	5. TM/TM Making or laughing about a joke
	4. Team/TM Speaking up with ideas	
<b>SPS ratio</b>		0.12
<b>5 longest lasting behaviours</b>		1. TM/TM Active listening
		2. TM/TM Making or laughing about a joke
		3. Team/TM Closed body language
		4. Team/TM Speaking up with ideas
		5. Team/TM Informing about issues or mistakes

Table 26: Comparison naked-eye and computer-aided observation + five longest lasting behaviours

## METHODOLOGICAL FINDINGS

### APPLICABILITY OF OBSERVATION SCHEME FOR COMPUTER-AIDED CODING

After three rounds of testing and coding one video fully the agreement between the researchers was 26.69% for the 20 minutes that have been turned into a golden file, and 31.49% for the whole video. The great difference in these two figures probably comes from the fact that the beginning of the meeting was rather chaotic leading to a lot of disparity in codes while the meeting was more tranquil towards the end leading to more agreement in coding.

Comparing the PS ratios that have been found (see Table 26), the two PS ratios based on counts are almost identical. This shows that computer-aided observation might not necessarily provide different or more accurate results than naked eye coding when looking at the global results. Moreover, looking at the PS ratio when calculated using the duration of behaviour, even this ratio is close to the ratio of the naked-eye observation based on counts. This further supports the similarity of naked-eye and computer-aided coding when used for global assessment of psychological safety.

Looking at the experience the two researchers had during coding with The Observer XT, the following conclusions can be made:

During coding it was sometimes difficult to correctly assess whether behaviour was directed at one individual or at the team as a whole. Complicating matters, sometimes behaviour was directed at two or three of the five participants which means that none of the two directions fit completely. This was also recognizable when comparing the results of the two researchers: On several occasions, the researchers had chosen the same behaviour but different directions of the behaviour. For the reliability analysis, such discrepancies were seen as complete disagreement as each behaviour has two different codes to account for two different directions. It is not possible to run a reliability analysis which considers only the behaviour irrespective of the direction.

Moreover, there was still some ambiguity regarding some of the behaviour codes. For example, it was never decided at which point in time the code *Reacting cold/ignoring a joke* should be coded. Should this be at the moment the joke is made, after the joke has been made or should it be for as long as the person makes the joke or for as long as the others are laughing about the joke? This also resulted in a lot of differences in coding, especially since the squad that was observed engaged a lot in joke-making. During comparison of the coding it was decided to place the code for *Reacting cold/ignoring a joke* as a point event directly after the speaker has finished making the joke.

Furthermore, there were still several examples of behaviour where both researchers didn't know exactly what behaviour they should code as the behaviour did not fit to any one category exactly. This was especially true for utterances such as "I don't know", "I don't care", "It doesn't matter to me".

Lastly, there was still some uncertainty regarding closed body language, where one of the researchers was unsure if putting one's hands before one's face or mouth would also count as closed body language. This led to a lot of differences in coding as well. In the end, it was decided to indeed include this behaviour as closed body language.

These kind of differences in interpretation of behaviours should not be possible in a well-developed observation scheme. Therefore, the observation scheme would still need some refinement before it is completely applicable.

---

#### NAKED-EYE OBSERVATION VERSUS COMPUTER-AIDED OBSERVATION

The researcher that engaged in both types of observation has made some qualitative observations about the two methods. More specifically, the advantages and disadvantages of computer-aided observation are elaborated on.

First of all, an advantage of computer-aided observation was that videos can be paused, rewind, slowed or re-watched without impairing reliability of the results. During naked-eye observation part of the shown behaviour was lost due to the researcher's inability to watch all people equally rigorously during the whole time of the video. Re-watching (parts of) some

videos would have then resulted in some videos being watched more closely than others, so this was not possible. During computer-aided observation the goal was that all behaviour was coded and, therefore, rewinding and re-watching the video was not only possible but also essential. The researchers watched the same video frame over and over until they were completely sure of the behaviour that should be coded. This also allowed for observing behaviours that were otherwise hard to code consistently, such as in this observation scheme, e.g. the *Use of inclusive language*, i.e. using the words “we”, “us” and “our”.

Secondly, during computer-aided observations, the data was instantly digitalized and could be exported as Excel files, removing the need to transform paper codes into digitalized codes. The Observer XT could even perform some analyses within the computer programme, such as calculating inter-rater reliability. This also led to easier comparison of reliability between observers, allowing even for intermediary analysis when the researchers were amidst observation. This could uncover inconsistencies in coding before the whole video is coded, and thus save time.

Lastly, observations using computer-programmes, such as The Observer XT, allow the researcher to capture more information on the behaviour displayed: The performer of the behaviour and duration of the behaviour can be coded as well.

Capturing who engaged in which behaviour could reveal individual differences between team members and their feelings of psychological safety. However, this was not particularly necessary in research that considers psychological safety at the team-level.

Duration of behaviour can be just as important as the times a behaviour as occurred. For example, when considering *Closed body language*, it could be possible that a person crosses their arms for a few seconds and then releases them, or a person crosses their arms and remains in this position for the rest of the meeting. When considering only counts of behaviour, one would say in both cases that only one instance of closed body language occurred which does not seem much. However, when considering the duration of behaviour as well, it can be seen that in the second case the person was engaging in closed body language for the whole of the meeting. This displays a critical difference.

However, the increased detail that can be captured using computer-aided observation comes at a cost: the time necessary to conduct observations. The naked-eye observation took as long as the meeting was, plus around two times the meeting length to read the transcript, and around 5-10 minutes to convert the scores that were made on paper into digital scores in Microsoft Excel. This means that observing 5 minutes of a meeting would take around 20-25 minutes. Using The Observer XT, observing the meeting took much longer, as much more detail was added to the analysis: it was not only analysed how often behaviour occurred but also how long and who was engaging in it. Additionally, the results had to be compared with a second researcher and a golden file needed to be made that combines both results. Observation in The Observer took around 22.5 hours for the whole video, meaning around 3 hours for 5 minutes. This is already much more but this does not include making the golden

file yet. This took another 5 hours per 5 minutes. So, in total, observation in The Observer XT takes 8 hours for 5 minutes of video. This is 19 times longer than the naked-eye observation.

Advantages	Disadvantages
Video can be paused, rewind, slowed and re-watched as often as needed without impairing reliability	Costs a lot of time, especially with a codebook as long as the one used in this research
Data is digitalized instantly, no need to transform counts on paper to Excel or SPSS	Always needs recording of meetings, so cannot be done by e.g. practitioners in real life
Allows for consistent observation of behaviours that are otherwise hard to observe (e.g. Silence behaviours)	
Considers duration of behaviour	
Can capture individual differences in behaviour	
Observations of different researchers can be easily compared	

*Table 27: Advantages and disadvantages of computer-aided observation*

In conclusion, the researcher experienced benefits and downsides of using one method over the other. These were mainly related to time necessary to conduct the observations versus quality and captured detail during the observation.

#### CROSS-STUDY COMPARISON

The cross-study comparison is structured as follows: First, the mean psychological safety for each study is compared to provide an overview. Then, the researcher compares the statistical findings on specific behaviours from Study 2 and 3 to see where they agree, where they disagree and where they add to each other. Next, the researcher compares the key findings from the triangulation of Study 1-3 and the results of Study 4. Finally, the researcher discusses which behaviours of the observation scheme have been hardly used or not at all in the naked-eye, as well as in the computer-aided observations.

Table 28 shows the mean and standard deviation of survey measured team psychological safety for each study. On average, the teams in Study 3 have by far the highest psychological safety.

	Mean	SD
<b>Study 1</b>	4.643	.492
<b>Study 2</b>	5.024	.698
<b>Study 3</b>	5.715	.497

*Table 28: Mean and standard deviation of team psychological safety per study*

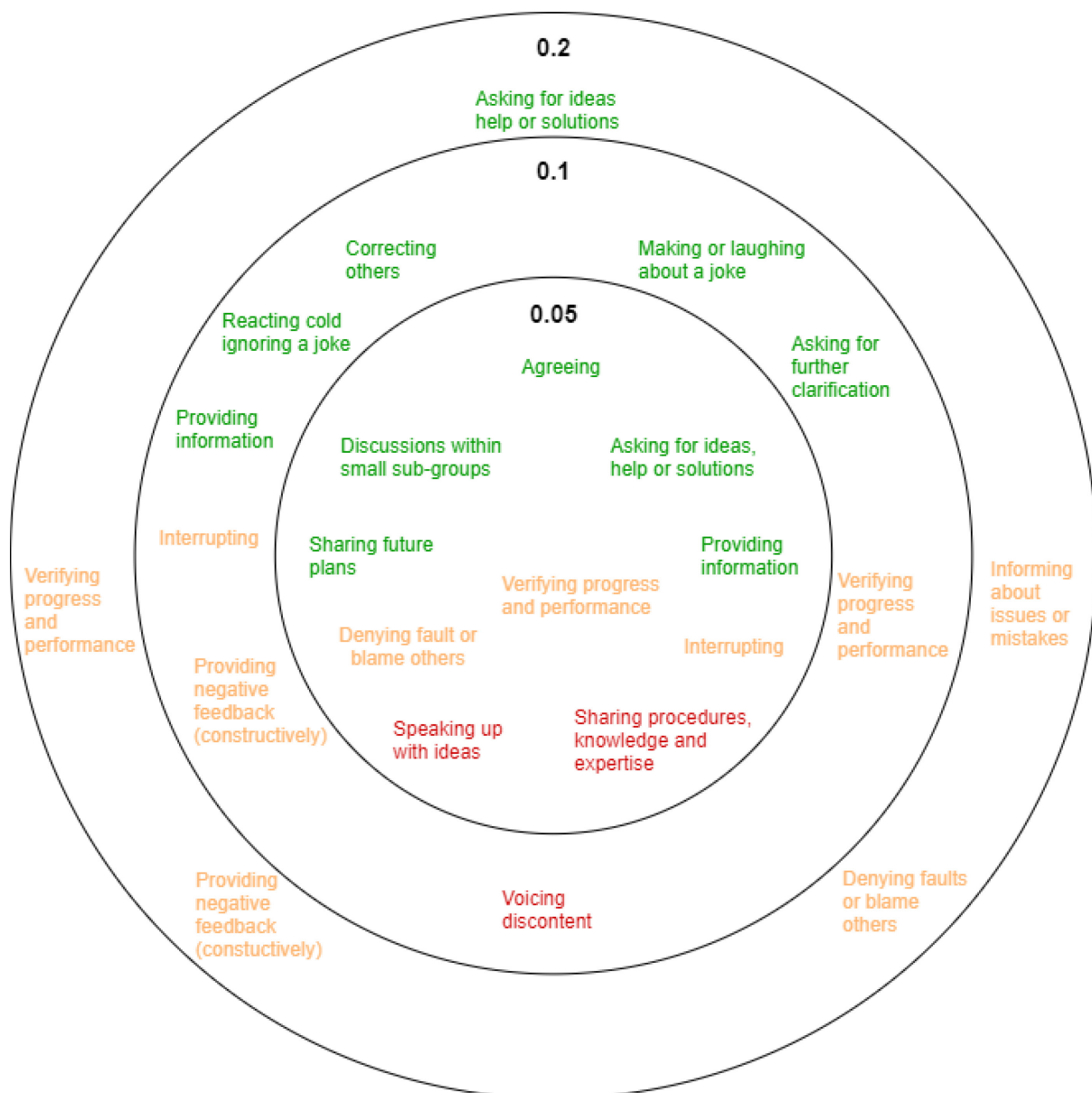


RELEVANT RELATIONSHIPS BETWEEN SPECIFIC BEHAVIOURS AND PSYCHOLOGICAL SAFETY

In total, for 17 of the 35 behaviours of the observation scheme, relevant relationships were found with psychological safety ( $p < 0.2$ ). Figure 1 summarizes which behaviours have been found to be related to psychological safety across the studies.

It can be seen that there have been many behaviours identified that have a positive relationship with psychological safety and only a couple that have a negative relationship. The yellow-coloured behaviours have been found to be related to psychological safety in both directions, so no clear conclusions can be made about these behaviours.

Figure 1: Summary of behaviours related to psychological safety, their significance and their direction



---

#### SUMMARY OF TOP FIVE OBSERVED BEHAVIOURS

Next, the top five observed behaviours in the teams that score highest on psychological safety, according to the survey (above 5.5), are summarized. These are Team 5 and 7 from Study 2, and Squad 1, 2, 3, and 6 from Study 3.

With the exception of *Making or laughing about a joke*, *Reacting cold/ignoring a joke* and *Disagreeing*, all behaviours that are included in the top five observed behaviours are in this list for at least two of the identified teams. It makes sense, therefore, to take a closer look at these behaviours as they seem to be prevalent in teams with exceptionally high psychological safety. The behaviours that are included on at least two of the teams are:

1. Agreeing (6 teams)
2. Active listening (5 teams)
3. Sharing procedures, knowledge and experience (3 teams)
4. Providing information (2 teams)
5. Interrupting (2 teams)
6. Asking for further clarification (2 teams)
7. Speaking up with ideas (2 teams)

From this list, for only three of these behaviours conclusive statistical findings were derived as can be seen when comparing the list with Figure 1. Two of the behaviours even have a negative statistical relationship with psychological safety but are still included in the top five most observed behaviours in the teams with the highest psychological safety levels.

---

#### BEHAVIOURAL CODES THAT HAVE HARDLY BEEN USED

The following table shows which behaviours have not been used at all or only once during the whole research. This considers the actual counts of behaviour, not the standardized counts. The purpose of this is to see whether there are behaviours in the observation scheme that might actually be redundant. *Talking about personal, non-work matters* was not used in any of the studies, *Facial expression or body language indicates fear* was only used once in all studies, and *Asking a question (other)* was only used twice across all studies.

---

#### TEAM PSYCHOLOGICAL SAFETY AND TEAM PERFORMANCE

The three studies that include team performance have all found similar relationships between psychological safety and team performance. In all studies the relationship was positive and marginally significant, or approach marginal significance ( $p = .200$ ,  $.200$  and  $.208$ ).

## DISCUSSION

This research intended to find out what the added value of using observational research methods in survey-based team psychological safety research is. Three assumptions were made at the start of the research, these being:

- (1) The observable psychological safety related behaviour differs between teams depending on their level of psychological safety.
- (2) Video-based research technology, i.e. The Observer XT, can aid in reliably identifying when behaviours related to psychological safety occur and can enrich data collection.
- (3) Teams with higher psychological safety have higher survey-reported team performance.

The discussion first assesses the theoretical implications of the results on these assumptions and the research question and then the practical implications. Lastly, limitations and avenues for future research are discussed.

## THEORETICAL IMPLICATIONS

Overall, the findings support the use of the three methods (a survey and quantitative and qualitative observations) in concert. The research, therefore, responds to the call for use of various research methods to assess psychological safety (Newman et al., 2017). There has been some alignment found between the quantitative observations and the surveys. The qualitative observations served to illuminate why certain differences in psychological safety between teams exist but also in some occasions why the findings from the quantitative observations and the survey do not align. The three methods can be seen as cumulative in the following order: The quantitative observations enlighten the survey results by defining what behaviours exactly occur more or less in psychologically safe teams. The qualitative observations further deepen these findings by providing context to these behaviours. These observations were especially helpful when behaviour was found to have an unexpected effect to explain why this effect is also plausible.

## OBSERVABLE BEHAVIOUR AND PSYCHOLOGICAL SAFETY

Regarding the first assumption, the research supports the argument that there are behaviours that occur more in psychologically safe teams than in others. These behaviours and the direction of their relationship have already been depicted in Figure 1.

The alignment with the theorized relationships by O'Donovan et al. (2020) varies. It seems that some behaviours have a different relationship with psychological safety than expected. This means that the categorization made by O'Donovan et al. (2020) that was used during this study for calculating the PS ratio might not be entirely true.

For example, *Discussion within small sub-groups* was not only found to be negatively related to psychological safety but also positively. This contradicts the expectation of O'Donovan et al. (2020) and moreover, contradicts the findings by Hoenderdos (2013) who found a clear negative relationship between the talking in sub-groups and psychological safety. The idea is that while discussing with only part of the team may be good for the relationship between the people in the sub-group, other people may feel left out and not dare to speak up to the sub-group. However, it might be that the positive effect of discussing in a small sub-group for the people in that sub-group exceeds the negative effect this

behaviour has on people excluded from the group, or that the people not in the sub-group do not see is as a big issue.

Furthermore, not every behaviour allows for fixing what effect it will have no matter the circumstances: For some behaviours there seem to be underlying nuances that can change the effect of the behaviour. Additionally, it can depend on the role of the person at whom the behaviour is directed at how the behaviour relates to psychological safety. Both of these implications can be explained using the example of Interrupting.

*Interrupting* was often one of the top five observed behaviours, irrespective of the team's surveyed psychological safety level. A potential explanation for this can be found in the literature on interrupting behaviour. The literature sees interruptions in two ways: they can be co-operative or intrusive (Murata, 1994). Co-operative interruptions show interest in the speaker's topic and do not intend to change the subject or take the floor. Relating this to psychological safety, such co-operative interruptions should not necessarily have an adverse effect on team psychological safety. On the other hand, intrusive interruptions are defined as interruptions that intend to change the subject of the conversation or take over the floor from the current speaker (Murata, 1994). These kinds of interruptions could harm the psychological safety team members feel as they devalue the elaborations of the interrupted speaker. These intricate differences in nuances of interrupting behaviour are not included in the observation scheme, which can explain why no univocal results are found on this behaviour.

Moreover, while *Interrupting* between team members is negatively related to psychological safety, team members interrupting the team leader is positively related to psychological safety. These differences can come from a difference in whose perspective is taken: (1) One can look at how psychologically safe the person that engages in the behaviour is feeling; or (2) one can look at how psychologically safe the person that is receptor of the behaviour is feeling. The observation scheme was initially meant to be used considering the second perspective. The first perspective can explain the positive relationship of team members interrupting team leaders with team psychological safety. Potentially, a team member will only interrupt his leader when he is feeling psychologically safe. Thus, there is a positive relationship. The second perspective can explain the negative relationship of team members interrupting each other with team psychological safety, which was also the expected direction of interrupting in general. Team members that get interrupted on might feel like their opinions are not heard or are quickly dismissed and, therefore, feel less psychologically safe. This shows that (1) behaviour does not have to have the same implications for all people, and (2) it is important to recognize that there are two perspectives to be taken into account when deciding what implications, a certain behaviour will have.

Lastly, for observers it is difficult to reliably assess in which of the many directions behaviour occurs. While it is relatively easy to distinguish whether behaviour is directed at the team leader or not, it is hard to say whether behaviour is directed at one team member or the team as a whole. This becomes especially difficult when behaviour is directed at two

or three of the team members but not all of them. Because differences exist depending on the direction of behaviour, it is not advised to leave out the directions altogether but it can be proposed to reduce the number of directions from five to three: Team members directing behaviour at the team leader or at the team, and the team leader directing behaviour at the team. This should result in a more straightforward distinction without eliminating too much detail.

---

#### COMPUTER-AIDED OBSERVATION

Regarding the second assumption, implications are that it depends on the research whether computer-aided observation provides benefits over naked-eye observation. Looking back at the advantages and disadvantages formulated in the findings, some of them were already defined in the literature. These being the ability to pause, rewind, slow or re-watch videos (Christianson, 2018; Noldus et al., 2000) without impairing reliability of the results, and the ability to include duration of behaviour in analysis (Christianson, 2018; Foster, 2006). Moreover, the computer-aided observation allows for analysis of which team members engage in which behaviour. This can be useful for some research goals.

This research showed that when looking exclusively at the relative frequencies of behaviour observed, naked-eye and computer-aided observation lead to similar results. However, a major concern derived from the elaboration of the researcher is the immense difference in time necessary for computer-aided observations versus naked-eye observations. Thus, ultimately, the decision to use computer-aided over naked-eye observation seems to be a trade-off between time and detail of observation. For example, if researchers are only interested in the ratio of behaviour indicative of high or low psychological safety and thus in triangulating their survey scores, naked-eye observation might be sufficient. The comparison between the PS ratio of the naked-eye observation show almost no difference with the PS ratio of the computer-aided observation when using frequencies of behaviour. However, if researchers are more interested in the underlying behaviours, especially when these are non-verbal behaviours or they want to know the duration and performer of the behaviours as well, computer-aided observation will be more insightful. The researchers should then be aware that data collection will be very time-consuming.

---

#### TEAM PSYCHOLOGICAL SAFETY AND TEAM PERFORMANCE

Regarding the last assumption, contrary to previous research (Kostopoulos & Bozionelos, 2011; Newman et al., 2017), this research does not find a significant relationship between team performance and psychological safety. An example that could explain this finding was found in the qualitative observations of one of the teams in Study 2. The team has quite high psychological safety and was engaging in a lot of behaviour indicative of psychological safety, particularly *Active listening* and *Agreeing*. However, no situations were observed in which the team members challenged each other's ideas. This could indicate that there exists some pressure to accept all input and conform with the group. This phenomenon of pressure to conform rather than thinking critically within a group has been conceptualized as groupthink

(Robbins & Judge, 2016). Groupthink itself can have negative effects on the group's performance. Potentially, this concept can stand in the way of psychological safety being positively related to psychological safety.

#### PRACTICAL IMPLICATIONS

The research has found that when wanting to assess the overall level of psychological safety, the PS ratio that is calculated by the frequency of specific behaviours can provide similar results to survey measurement. Also, the exploration of computer-aided observation has revealed that naked-eye observations provide similar results when only considering the level of psychological safety measured (PS ratio), further supporting this use of the observation scheme for naked-eye assessments. Therefore, practitioners could use the observation scheme to assess the level of psychological safety in their teams with naked-eye observation. This could even be done repeatedly without exhausting the team members as they just follow their normal meetings.

Moreover, the findings of the observation scheme regarding specific behaviours that occurs more often in psychologically safe teams can inform practitioners on what behaviours to encourage when wanting to increase psychological safety. These behaviours are *Agreeing*, *Discussions within small sub-groups*, *Asking for ideas, help or solutions*, *Sharing future plans* and *Providing information*.

#### LIMITATIONS AND FUTURE RESEARCH

Limitations pertain to four components of the research: the quality of recordings, the sample, and, the observation scheme and the methodology in general.

Firstly, the quality of recordings differed substantially across and within the studies. Some recordings were made with very advanced equipment and from an angle that shows all participants clearly, such as the recordings of Study 3. In other recordings, some participants were left out of the video frame or the frame was mobile, meaning that there were no moments at which all participants were visible simultaneously. Also, there were recordings in which the quality of the sound was poor. All of these differences are detrimental to the comparability of results between the teams. Moreover, the lower quality of some of the recordings limited the researcher's ability to code the videos consistently which is detrimental to the results for these teams.

Future research should ensure that all meetings are recorded in a consistent way and from an angle at which all participants are visible. Moreover, the camera should be stable and the audio recordings should be clear.

Secondly, the sample in each study is very small, Dutch participants make up a majority of the sample, and there is a wide range in length of sampled meetings. The small sample size limits the reliability and generalizability of the findings of this research. Also, the large number of Dutch participants limits the generalizability to other cultures. It could already be seen that in the one team from Study 3, in which all participants reported the

category ‘Other’ as their nationality, very inconsistent findings were found. This could pertain both to these participants interpreting survey questions differently, and finding different behaviours appropriate in different contexts as well as the researcher interpreting the behaviours of these participants incorrectly. Cultural differences might deter the reliability of results. Lastly, the sample included meetings of various length also impacting the reliability of comparisons across teams. The effect hereof was attempted to be mitigated by standardizing the observation to a 10 minute and 5 people average but it can still be expected that the psychological safety for a team that was observed for 20 minutes is more accurate than the psychological safety for a team that was observed only 5 minutes.

Future research should try to recruit a larger sample within one study to allow for more reliable statistical and qualitative results. Ideally, this sample would encompass a range of cultural backgrounds, or several studies would compare samples from different cultural backgrounds, to test whether the observational scheme is valid in non-Dutch cultures as well. Moreover, future research should try to observe meetings of similar length. It can be expected that after watching a certain amount of time of one team, the value of additional observations becomes smaller. Thus, there would be an ideal length that should be observed. Future research could find out what that length is and adhere to watching that amount of time of all meetings in their sample.

Furthermore, the research has shown that the observation scheme is not yet optimal. There are some behaviours that were missing from the observation scheme while others were obsolete (see Appendix VII). This limits the reliability and validity of the findings that were made using this observation scheme.

Future research should reconsider the observation scheme and adapt it where necessary based on more empirical research.

Lastly, several methodological limitations can be identified. Firstly, the naked-eye observations in this study are conducted by only one researcher. Considering the low agreement that was achieved during the computer-aided observation between the two researchers, the results from the naked-eye observation might be questionable. Secondly, the qualitative observations in this study follow no pre-determined strategy or methodology but are done instinctively. Thirdly, the studies use different surveys to assess psychological safety, which impairs the cross-study comparability. Furthermore, some behaviours, such as the *Use of inclusive language* were identified to be difficult to observe in naked-eye observation. Moreover, the observations, especially the computer-aided observations were deemed very time-consuming, limiting the applicability of the method to larger samples. Additionally, the study is cross-sectional and thus no causal inferences can be made. It cannot be determined whether the behaviours elicit psychological safety or the behaviours are elicited by psychological safety. Lastly, the observation scheme does not include behaviours that actually measure Silence behaviour as it is conceptualized: Purposefully withholding relevant information from the group.

Future research should recruit at least two researchers to conduct the naked-eye observations and measure their agreement. As the naked-eye observations can be done with video-recorded meetings this poses no additional reactivity threats. Secondly, future research should determine a strategy for the qualitative observations of the meetings. Thirdly, future research should compare observational findings with one psychological safety survey. Specifically, the survey of O'Donovan et al. (2020) would be advised to use as this survey has been specifically developed for use in concert with the observation scheme. Furthermore, *Use of inclusive language* could be measured while reading the transcripts of the meetings by coding how often the words “we”, “us” and “our” occur. This, however, prerequisites that transcripts are made at all. Moreover, in the future, time necessary for observation could be diminished by use of Artificial Intelligence, at least for some of the well-defined, straightforward behaviours such as *Active listening* (only and always coded when looking at the speaker), *Use of inclusive language* (only coded when word “we”, “us” or “our” is used), or *Closed body language* (only coded when a person crosses his arms or puts hands in front of the face). This could already reduce the time necessary for observations to some degree. Furthermore, future research should aim to use the observational scheme in longitudinal research to define whether behaviour follows psychological safety or the other way around. Lastly, a method of data collection on true *Silence behaviour* should be included, so that the research actually encompasses all behaviours that pertain to psychological safety. This is especially important since research has found *Silence behaviours* to relate more strongly to psychological safety than *Voice behaviours* (Sherf et al., 2020). It has been found very hard to observe *Silence behaviours* as they are less obtrusive than *Voice behaviours* (Van Dyne et al., 2003) and it is difficult to determine whether people are silent because they are withholding information or simply because they have nothing to say (Meinecke et al., 2016). O'Donovan and McAuliffe (2020b) have included interviews with participants in their observational and survey research on psychological safety in the health sector. Including such interviews can also shed light on true *Silence behaviour* of participants by investigating whether participants actively withhold information and why. Thus, it would be advised to include this in future research.

## CONCLUSION

In conclusion, what is the added value of observational research methods in team psychological safety research? The main value provided by observational research methods is the verification (through triangulation) and enrichment (through crystallization) of survey results. The observation scheme allowed the researcher to find specific behaviours that are related to team psychological safety. Especially, a great number of behaviours that potentially have positive relationships with psychological safety have been found. However, a number of flaws of the current observation scheme have also been pointed out, particularly regarding the direction of the relationship with psychological safety, the multiple directions of behaviour that can be chosen, and the extent of the current observation scheme, indicating



that the scheme can still be improved. Future research should test the found relationships with larger samples to arrive at more robust results which can also inform what behaviours are essential to include in the observation scheme. Finally, the research has found that computer programmes, such as The Observer XT, can enrich observational analysis even further by capturing more detail, specifically through duration of behaviour and performer of behaviour.

## REFERENCES


- Annosi, M. C., Magnusson, M., Martini, A., & Appio, F. P. (2016). Social conduct, learning and innovation: An abductive study of the dark side of agile software development. *Creativity and Innovation Management, 25*(4), 515-535.
- Baer, M., & Frese, M. (2003). Innovation is not enough: climates for initiative and psychological safety, process innovations, and firm performance. *Journal of Organizational Behaviour, 24*, 45-68.
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396-403.
- Beckett, M., Da Vanzo, J., Sastry, N., Panis, C., & Peterson, C. (2001). The quality of retrospective data: An examination of long-term recall in a developing country. *The Journal of Human Resources, 36*(3), 593-625.
- Brady, F. N. (1985). A Janus-headed model of ethical theory: Looking two ways at business/society issues. *Academy of Management Journal, 10*, 568-576.
- Brinsfield, C. T. (2013). Employee silence motives: Investigation of dimensionality and development of measures. *Journal of Organizational Behaviour, 34*, 671-697.
- Castiglioni, L., Pforr, K., & Krieger, U. (2008). The effect of incentives on response rates and panel attrition: Results of a controlled experiment. *Survey Research Methods, 2*(3), 151-158.
- Christianson, M. K. (2018). Mapping the terrain: The use of video-based research in top-tier organizational journals. *Organizational Research Methods, 21*(2), 261-287.
- Delgado Pina, M. I., Romero Martinez, A. M., & Gomez Martinez, L. (2008). Teams in organizations: a review on team effectiveness. *Team Performance Management, 14*(1/2), 7-21.
- Detert, J. R., & Burris, E. R. (2007). Leadership behavior and employee voice: Is the door really open? *Academy of Management Journal, 50*(4), 869-884.
- Detert, J. R., Burris, E. R., Harrison, D. A., & Martin, S. (2013). Voice flows to and around leaders: Understanding when units are helped or hurt by employee voice. *Administrative Science Quarterly, 58*(4), 624-668.
- Detert, J. R., & Edmondson, A. C. (2011). Implicit voice theories: Taken-for-granted rules of self-censorship at work. *Academy of Management Journal, 54*(3), 461-488.
- Donaldson, S. I., & Grant-Vallon, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2), 245-260.
- Dooley, D. (2009a). Inferential statistics: Drawing valid conclusions from samples. In *Social Research Methods* (4 ed., pp. 143-160). New Jersey: Pearson Education.
- Dooley, D. (2009b). Survey data collection: Issues and methods in sample surveys. In *Social Research Methods* (4 ed., pp. 117-142). New Jersey: Pearson Education.
- Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*, 350-383.
- Edmondson, A. C., & Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal construct. *The Annual Review of Organizational Psychology and Organizational Behaviour, 1*, 23-43.
- Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *Academy of Management Review, 32*(4), 1155-1179.

- Foster, P. (2006). Observational Research. In R. Sapsford & V. Jupp (Eds.), *Data Collection and Analysis* (2 ed., pp. 57-92). London: SAGE.
- Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pzeshkan, A., & Vracheva, V. (2017). Psychological safety: A meta-analytic review and extension. *Personnel Psychology, 70*, 113-165.
- Gibson, C. B., Cooper, C. D., & Conger, J. A. (2009). Do you see what we see? The complex effects of perceptual distance between leaders and teams. *Journal of Applied Psychology, 94*(1), 62-76.
- Hill, A. D., White, M. A., & Wallace, J. C. (2014). Unobtrusive measurement of psychological constructs in organizational research. *Organizational Psychology Review, 4*(2), 148-174.
- Hoenderdos, J. W. (2013). *Towards an observational measure for team psychological safety*. (Master Thesis). University of Twente, The Netherlands.
- Hoogeboom, M. A. M. G., & Wilderom, C. P. M. (2020). A complex adaptive systems approach to real-life team interaction patterns, task context, information sharing, and effectiveness. *Group & Organization Management, 45*(1), 3-42.
- Kahn, W. A. (1990). Psychological conditions of personal engagement and disengagement at work. *Academy of Management Journal, 33*(4), 692-724.
- Kim, J. H., & Choi, I. (2019). Choosing the level of significance: A decision-theoretic approach. *ABACUS, 1-45*. doi:10.1111/abac.12172
- Klaas, B. S., Olson-Buchanan, J. B., & Ward, A. (2012). The determinants of alternative forms of workplace voice: An integrative perspective. *Journal of Management, 38*(1), 314-345.
- Klonek, F. E., Gerpott, F. H., Lehmann-Willenbrock, N., & Parker, S. K. (2019). Time to go wild: how to conceptualize and measure process dynamics in real teams with high-resolution. *Organizational Psychology Review, 9*(4), 245-275.
- Kostopoulos, K. C., & Bozionelos, N. (2011). Team exploratory and exploitative learning: Psychological safety, task conflict, and team performance. *Group & Organization Management, 36*(3), 385-415.
- Kozlowski, S. W. J. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review, 5*(4), 270-299.
- Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*(3), 77-124.
- LeBaron, C., Jarzabkowski, P., Pratt, M. G., & Fetzer, G. (2018). An introduction to video methods in organizational research. *Organizational Research Methods, 21*(2), 239-260.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852.
- LePine, J. A., & Van Dyne, L. (1998). Predicting voice behaviour in work groups. *Journal of Applied Psychology, 83*(6), 853-868.
- Liang, J., Farh, C. I. C., & Farh, J. L. (2012). Psychological antecedents of promotive and prohibitive voice: A two-wave examination. *Academy of Management Journal, 55*, 71-92.
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Challenging-oriented organizational citizenship behaviors and organizational effectiveness: Do challenge-oriented behaviors really have an impact on the organization's bottom line? *Personnel Psychology, 64*(3), 559-592.

- Martins, L. L., Schilpzand, M. C., Kirkman, B. L., Ivanaj, S., & Ivanaj, V. (2013). A contingency view of the effects of cognitive diversity on team performance: The moderating roles of team psychological safety and relationship conflict. *Small Group Research, 44*(2), 96-126.
- Mathieu, J. E., Hollenbeck, J. R., Van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of Applied Psychology, 102*(3), 452-467.
- Meinecke, A. L., Klonek, F. E., & Kauffeld, S. (2016). Using observational research methods to study voice and silence in organizations. *German Journal of Human Resource Management, 30*(3-4), 195-224.
- Milliken, F. J., Morrison, E. W., & Hewlin, P. F. (2003). An exploratory study of employee silence: Issues that employees don't communicate upward and why. *Journal of Management Studies, 40*(6), 1453-1476.
- Morrison, E. W. (2014). Employee voice and silence. *The Annual Review of Organizational Psychology and Organizational Behaviour, 1*, 173-197.
- Morrison, E. W., & Milliken, F. J. (2000). Organizational silence: a barrier to change and development in a pluralistic world. *Academy of Management Review, 25*, 706-725.
- Murata, K. (1994). Intrusive or co-operative? A cross-cultural study of interruption. *Journal of Pragmatics, 21*, 385-400.
- Nembhard, I. M., & Edmondson, A. C. (2006). Making it safe: The effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *Journal of Organizational Behaviour, 27*, 941-966.
- Nemeth, C. J., Connell, J. B., Rogers, J. D., & Brown, K. S. (2006). Improving decision making by means of dissent. *Journal of Applied Psychology, 31*(1), 48-58.
- Newman, A., Donohue, R., & Eva, N. (2017). Psychological safety: A systematic review of the literature. *Human Resource Management Review, 27*, 521-535.
- Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The observer video-pro: New software for the collection, management and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, 32*(1), 197-206.
- O'Donovan, R., & McAuliffe, E. (2020a). A systematic review exploring the content and outcomes of interventions to improve psychological safety, speaking up and voice behaviour. *BMC Health Services Research, 20*(101).
- O'Donovan, R., & McAuliffe, E. (2020b). Exploring psychological safety in healthcare teams to inform the development of interventions: combining observational, survey and interview data. *BMC Health Services Research, 20*(810).
- O'Donovan, R., Van Dun, D., & McAuliffe, E. (2020). Measuring psychological safety in healthcare teams: Developing an observational measure to complement survey methods. *BMC Medical Research Methodology, 20*(203).
- Pearsall, M. J., & Ellis, A. P. J. (2011). Thick as thieves: The effect of ethical orientation and psychological safety on unethical team behavior. *Journal of Applied Psychology, 96*(2), 401-411.
- Penner, L. A., Orom, H., Albrecht, T. L., Franks, M. M., Foster, T. S., & Ruckdeschel, J. C. (2007). Camera-related behaviours during video recorded medical interactions. *Journal of Nonverbal Behaviour, 31*, 99-117.

- Pinder, C. C., & Harlos, K. P. (2001). Employee silence: Quiescence and Acquiescence as Responses to Perceived Injustice. *Research in Personnel and Human Resources Management, 20*, 331-369.
- Ployhart, R. E., & Ward, A. (2011). The "Quick Start Guide" for conducting and publishing longitudinal research. *Journal of Business and Psychology, 26*, 413-422.
- Pugliese, A., Nicholson, G., & Bezemer, P. (2015). An observational analysis of the impact of board dynamics and directors' participation on perceived board effectiveness. *British Journal of Management, 26*, 1-25.
- Robbins, S. P., & Judge, T. A. (2016). Key group concepts. In *Essentials of Organizational Behaviour* (13 ed., pp. 196-212). England: Pearson Education Limited.
- Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: discoveries and developments. *Human Factors, 50*(3), 540-547.
- Sherf, E. N., Parke, M. R., & Isaakyan, S. (2020). Distinguishing voice and silence at work: Unique relationships with perceived impact, psychological safety, and burnout. *Academy of Management Journal*.
- Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 2*(1), 16-18.
- Stray, V., Sjøberg, S. I. K., & Dybå, T. (2016). The daily stand-up meeting: A grounded theory study. *The Journal of Systems and Software, 114*, 101-124.
- Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837-851.
- Van Den Bossche, P., Gijssels, W. H., Segers, M., & Kirschner, P. A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors. *Small Group Research, 37*(5), 490-521.
- Van Dyne, L., Ang, S., & Botero, I. C. (2003). Conceptualizing employee silence and employee voice as multidimensional constructs. *Journal of Management Studies, 40*(6), 1359-1392.
- Verhelst, F. (n.d.). Weekly stand-up uitleg en tips. Retrieved from <https://agilescrumgroup.nl/weekly-stand-up/>
- Waller, M. J., & Kaplan, S. A. (2018). Systematic behavioral observation for emergent team phenomena: Key considerations for quantitative video-based approaches. *Organizational Research Methods, 21*(2), 500-515.
- Walumbwa, F. O., & Schaubroeck, J. (2009). Leader personality traits and employee voice behavior: Mediating roles of ethical leadership and work group psychological safety. *Journal of Applied Psychology, 94*(5), 1275-1286.
- Wang, M., Beal, D. J., Chan, D., Newman, D. A., Vancouver, J. B., & Vandenberg, R. J. (2017). Longitudinal research: A panel discussion on conceptual issues, research design and statistical techniques. *Work, Aging and Retirement, 3*(1), 1-24.
- Weiss, M., Kolbe, M., Grote, G., Dambach, M., Marty, A., Spahn, D. R., & Grande, B. (2014). Agency and communion predict speaking up in acute care teams. *Small Group Research, 45*(3), 290-313.
- Weiss, M., Kolbe, M., Grote, G., Spahn, D. R., & Grande, B. (2018). We can do it! Inclusive leader language promotes voice behavior in multi-professional teams. *The Leadership Quarterly, 29*, 389-402.

Zhang, X., Liang, L., Tian, G., & Tian, Y. (2020). Heroes or villains? The dark side of charismatic leadership and unethical pro-organizational behavior. *International Journal of Environmental Research and Public Health*, 17(15), 1-16.



## APPENDIX

## APPENDIX I – THE ISSUE OF NORMALITY

Sample sizes for all studies in this research were very small, with the largest study concerning only 7 teams. Under such low sample sizes, it becomes impossible to reliably assess normality. For this, it would be recommended to have samples of at least 25 teams. Subsequently, only non-parametric measures were used throughout the whole research as normality cannot be assumed.

## APPENDIX II – SIGNIFICANCE LEVELS

While this research did not focus on hypothesis testing due to the small sample sizes in each study, correlation tests did form part of the research, making it necessary to consider significance levels.

The level of significance is the probability that one rejects the null hypothesis while it is actually true (Skipper, Guenther, & Nass, 1967). This is the same as the probability of making a Type I error. Consequently, the level of significance is also closely related to the probability of making a Type II error: accepting the null hypothesis while it false. Typically, researchers set a threshold as to which significance levels are appropriate and call results that fall below that level 'significant'. The most commonly used significance level is 0.05 (Dooley, 2009a; Skipper et al., 1967). Using this significance level, there is a chance of 5% that the results found in the research are actually false. However, the blind usage of this level in every type of research has been critiqued for being too arbitrary (Kim & Choi, 2019; Skipper et al., 1967). Instead, the specific research problem should determine what kind of significance level is appropriate as different problems allow for different levels of probability for Type I and Type II errors. Additionally, the significance level should be in accordance with the sample size. In some situations, making a Type I error would be more consequential while in others making a Type II error would be more consequential (Skipper et al., 1967). Additionally, the significance level should be in accordance with the sample size. If the significance level is fixed at 0.05, the power of the test decreases with a smaller sample size (Kim & Choi, 2019). As the studies in this research all have very small sample sizes, this is the main point to consider. The minimum significance level used for this research was 0.2, meaning that there is a 20% chance that the results are not true. Results that fall under this level can be called marginally significant.

APPENDIX III – ORIGINAL OBSERVATION SCHEME (O'DONOVAN ET AL., 2020)

	Psychological Safety Towards Team Leader	Psychological Safety Towards Other Team Members		Psychological Safety in Relation to Team as a Whole	
	Team Members	Team Leader	Team Members	Team Leader	Team Members
<b>VOICE BEHAVIOURS</b>					
Communicating opinions to others even if they disagree					
Asking questions					
Providing information					
Providing feedback					
Providing help or solutions					
Correcting others					
<b>DEFENSIVE VOICE BEHAVIOURS</b>					
Denying faults or blame others					
Showing aggression					
Evading confrontation by focusing only on positives					
<b>SILENCE BEHAVIOURS</b>					
Facial expression or body language indicates fear					
Facial expression or body language indicates disengagement					
Closed body language					
<b>SUPPORTIVE BEHAVIOURS</b>					



Sharing procedures, knowledge and experience					
Sharing future plans					
Active listening					
Use of inclusive language such as “we”					
Agreeing/Responding positively or enthusiastically to input					
Acknowledging achievements/ congratulating					
Delegating tasks					
<b>UNSUPPORTIVE BEHAVIOURS</b>					
Interrupting					
Discussions within small sub-groups					
Reacting cold/ignoring a joke					
<b>LEARNING OR IMPROVEMENT ORIENTED BEHAVIOURS</b>					
Reviewing own progress and performance					
Asking for feedback					
Asking for help or solutions					
Asking for input from all meeting participants					
Informing the team about issues or mistakes related to patient safety or staff safety					
Looking for improvement opportunities and speaking up with ideas					
Acknowledging own mistake					
<b>FAMILIARITY BEHAVIOURS</b>					

Talking about personal, non-work matters					
Laughing about a joke					
<b>TOTAL OBSERVED BEHAVIOUR</b>					
<i>Categories indicating high psychological safety: (voice behaviours, supportive behaviours, learning or improvement behaviour and familiarity behaviours)</i>					
<i>Categories indicating lower psychological safety: (defensive voice behaviours, silence behaviours and unsupportive behaviours).</i>					

---

 APPENDIX IV – ADAPTATIONS TO THE OBSERVATION SCHEME
 

---

 CHANGES AFTER NAKED-EYE OBSERVATION
 

---

1. *Asking for further clarification*: This code was added under the category Voice Behaviour to account for behaviour in which team members asked for repetition of what was just said or for more in-depth elaborations on the specific topic.
2. *Asking a question (other)*: This code was added under the category Voice Behaviour as it was sometimes difficult to categorize questions ad hoc. The researcher could then use this code to show that a question was raised if the researcher didn't know to which category it would fit.
3. *Providing feedback*: This code was split into two codes, providing negative and providing positive feedback. While both codes were seen as Voice Behaviour, thus indicative of high psychological safety, this enabled the researcher to discover more nuance in the behaviour of the participants.
4. *Voicing discontent*: This code was added under the category Voice Behaviour for situations in which people were complaining about situations or people outside the team. This behaviour was not always about a concrete issue or mistake so it did not fit into the category Informing about issues or mistakes.
5. *Acknowledging achievements/congratulating*: The behaviour of "Thanking" was added to the list of examples.
6. *Accepting feedback*: This code was added under the category Learning and improvement behaviour as this could show how people reacted to feedback. This kind of behaviour could be related to a higher level of team psychological safety.
7. *Informing about issues or mistakes*: The part "related to patient safety or staff safety" was omitted because it was not relevant in the various business contexts in this research.

 CHANGES AFTER ROUND 1 COMPUTER-AIDED OBSERVATION
 

---

First of all, a change was made in the methodological approach to the observations. It was agreed to observe Active Listening and Silence behaviours separately to allow full focus on these behaviours. Secondly, it was decided that each utterance could only have one code, e.g. Providing information and Sharing procedures, knowledge and experience could not be coded simultaneously. A choice needed to be made to which code fits the behaviour best. Additionally, the following changes have been made to the observation scheme:

1. *Communicating opinions to others even if they disagree*: This name of this code was not in line with its definition and examples. Consequently, it became ambiguous when this code should be used. It was decided to align the naming of the code with the definition and examples by calling it *Disagreeing*.
2. *Providing information*: This code was unclear and was often coded simultaneously with other codes in which some form of information was given. The idea, however,

was to have each behaviour be assigned only one code, so this was not allowed. The definition of the code was refined to include that it is only about factual information. This should resolve the ambiguity around this code.

3. *Providing negative feedback*: In the definition of this code it said “constructively” which is not embedded in the name of this code. This led to researchers interpreting it differently. To make this clearer the code was altered to *Providing negative feedback (constructively)*.
4. *Showing aggression*: There was some mocking behaviour observed in the test round and the researchers did not know how to code such behaviour. It was decided that it could count as a micro-aggression and should, therefore, be coded under *Showing aggression*. The word micro was added in the definition of this code.
5. *Use of inclusive language*: There was some discrepancy in how researchers interpreted this code. While one researcher focused on the word “we” and used the code every time someone used this word, the other researcher looked at contexts and used this code when a person was, for example, talking about how the team together could solve a problem. Ultimately, it was decided to use this code whenever someone used the words “we”, “us”, or “our”. It was then also decided that it is a point event, so for this code no duration was measured.
6. *Interrupting*: It was ambiguous for how long this behaviour should be coded. It was decided to set this behaviour as a point event, as well, at the moment of interruption. So, also for this code no duration was measured.
7. *Reviewing progress and performance*: This code was misleading as the researchers interpreted it to be used whenever the team was talking about their progress or performance. However, the researcher that developed the observation scheme, clarified that it was intended to be used only when a question is raised about progress and performance. To make this clearer, the name of the code was altered towards *Verifying progress and performance*.
8. *Asking for input from all meeting participants*: This code was omitted and included in the code *Asking for help or solutions*. The new code is *Asking for ideas, help or solutions*. It is not necessary to determine the direction of the code in the name of the code as this can be done via the various columns. That is why the part “from all meeting participants” is deleted.
9. *Looking for improvement opportunities and speaking up with ideas*: This code actually included two different behaviours. It was, therefore, decided to shorten it and was named only *Speaking up with ideas*. *Looking for improvement opportunities* was not added elsewhere in the observation scheme.

---

#### CHANGES AFTER ROUND 2 COMPUTER-AIDED OBSERVATION

1. *Providing negative feedback (destructively)*: This code was added under *Defensive voice behaviour* to account for negative feedback that was not given in a nice way. It was expected that this would be detrimental to psychological safety.

2. *Denying fault or blaming others*: Clarified that these are two independent behaviours and when someone is blaming another without denying fault, this should also be coded.
3. *Facial expression or body language indicates disengagement*: When a person quickly looks on his watch or phone to find out what time it is, this is not seen as disengagement.
4. *Sharing procedures, knowledge and experience*: This code is also used when someone is warning the team about potential consequences of actions. It is interpreted that warning is based on this person's experience with these actions.
5. *Sharing future plans*: Also sharing of visions for the future is coded as Sharing future plans. This means that the plan does not have to be set in stone yet.
6. *Active listening*: Clarified that this code is only used when a person looks at the speaker for at least 1 second. Whenever a person engages in different behaviour, active listening is stopped. Even if it is just a short utterance. Lastly, it is decided that active listening is only coded in the individual direction, as it is only possible to actively listen to one person at a time.
7. *Use of inclusive language*: When a person is quoting what people outside the team said, and in these quotes the words "we", "us", or "our" is used, this is not coded as inclusive language. The word "Let's" is only coded as inclusive language when it is used in an inclusive way. For instance, when someone says "Let's do this together!" this would be inclusive language but when someone says "Let's see" this is not seen as inclusive language.
8. *Laughing about a joke*: To this code Making a joke is added.
9. *Making a negative joke*: This behaviour is added to the Unsupportive behaviours to account for sarcasm or cynicism.

Lastly, a checklist was made to include all the steps that should be followed when using the observation scheme for psychological safety in The Observer XT. This was done in Dutch because both researchers that were observing the meetings were Dutch. The checklist can be found in Appendix II.

---

#### CHANGES AFTER ROUND 3 COMPUTER-AIDED OBSERVATION

1. *Disagreeing*: This code is only used in the individual direction, as one always disagrees with a statement a person made. Moreover, it was unclear when to use Disagreeing and when to use Correcting others. It was decided that these codes are often used in sync. The first part of a sentence would be Disagreeing and the second part Correcting others. For example: No (= Disagreeing), you should do it this way (= Correcting others).
2. *Asking a question (other)*: This code was not used at all during observation in The Observer XT. It was decided that if it was needed, the researchers write down for what question it was used to see if it really didn't fit with another code. Potentially, this code could be scrapped if the researchers still didn't use it.

3. *Silence behaviour*: It was agreed that all silence behaviours could only be coded towards the team as a whole.
4. *Closed body language*: This code is only about the upper body, so crossing one's legs, for example, is not included. The example picture that showed this behaviour in the observation scheme was deleted.
5. *Active listening versus Agreeing*: As both behaviours included some form of nodding it was sometimes hard to distinguish between the two. It was agreed that to differentiate between them the context should be taken into account. When the previous utterance was something one could agree (or disagree) to this code would be used. If this was not the case, the nodding was seen as neutral and thus as Active listening.
6. *Agreeing*: Similar to Disagreeing, this behaviour is always directed at an individual team member.
7. *Reacting cold/ignoring a joke*: This code does not mean that someone is purposefully ignoring a joke. It should always be coded when someone does not react to a joke, thus when a joke is made, every team member gets a code: either Reacting cold/ignoring a joke or Making or laughing about a joke.
8. *Making a negative joke*: This code was deleted again and behaviour that could fall under this category should be placed under Providing negative feedback (destructively).
9. *Informing the team about issues or mistakes*: The name of this code was altered to be Informing about issues or mistakes as the direction of the behaviour could be chosen in the columns of the scheme.
10. *Making or laughing about a joke*: Making a joke is always coded towards the whole team, while laughing about a joke is always coded as a behaviour towards one individual. Consequently, also reacting cold/ignoring a joke is only coded towards one individual.

---

#### APPENDIX V – CHECKLIST FOR OBSERVATION IN THE OBSERVER XT

1. Word-bestand "Video-Observation Scheme" printen en de bijbehorende toetsen uit The Observer per gedrag invullen voor een makkelijk overzicht tijdens het coderen
2. Checken welk nummer elke deelnemer heeft
3. **Een voor een** de deelnemers over de gehele video observeren en **uitsluitend** op de volgende codes letten:
  - Facial expression indicates fear
  - Facial expression or body language indicates disengagement
  - Closed body language
  - Active listening
4. Nu de **hele video opnieuw kijken** en alle verbale gedragingen coderen (met uitzondering van "inclusive language"). Daarbij opletten:
  - Als iemand een grap maakt, moeten alle deelnemers een code krijgen. Het is of "making or laughing about a joke" of "reacting cold/ignoring a joke". **Iedereen** moet

een van deze twee codes krijgen als er een grap wordt gemaakt. Ook als het lijkt alsof iemand de grap gewoon niet heeft gehoord.

- Active listening **moet altijd uitgezet worden** als iemand aan het praten is of met een andere verbale gedraging bezig is (hierbij telt bijvoorbeeld agreeing door ja knikken met het hoofd ook als verbale gedraging)
5. Checken of “active listening” altijd uit staat als een verbaal gedrag wordt vertoond. (Een manier om dit makkelijker te doen:
- Je bestand opnieuw opslaan onder een nieuwe naam.
  - Daarvan een .odx file maken.
  - Een Reliability Analysis starten met je eigen odx.file.
  - Deze filteren per follower
  - In de Comparison List kun je nu makkelijk per follower zien wanneer je Active Listening aan hebt laten staan terwijl er ander gedrag vertoond werd.
  - Je kunt je echte file tegelijkertijd openen en daarin de observatie aanpassen.
6. De hele video opnieuw doornemen en elk moment dat iemand “we”, “us” of “our” zegt, coderen als inclusive language.

APPENDIX VI – ADAPTED OBSERVATION SCHEME

	Psychological Safety Towards Team Leader	Psychological Safety Towards Individual Team Members		Psychological Safety in Relation to Team as a Whole	
	Team Members	Team Leader	Team Members	Team Leader	Team Members
<b>Voice Behaviours</b>					
Disagreeing				-----	-----
Asking for further clarification					
Asking a question (other)					
Providing information					
Providing positive feedback					
Providing negative feedback (constructively)					
Providing help or solutions					
Correcting others					
Voicing discontent					
<b>Defensive Voice Behaviours</b>					
Providing negative feedback (destructively)					
Denying faults or blame others					
Evading confrontation					
Showing aggression					
<b>Silence Behaviours</b>					
Facial expression or body language indicates fear	-----	-----	-----		



Facial expression or body language indicates disengagement	-----	-----	-----		
Closed body language	-----	-----	-----		
<b>Supportive Behaviours</b>					
Sharing procedures, knowledge and experience					
Sharing future plans					
Active listening				-----	-----
Agreeing/Responding positively or enthusiastically to input				-----	-----
Acknowledging achievements/ congratulating					
Delegating tasks					
Use of inclusive language					
<b>Unsupportive Behaviours</b>					
Interrupting					
Discussions within small sub-groups				-----	-----
Reacting cold/ignoring a joke				-----	-----
<b>Learning or Improvement Oriented Behaviours</b>					
Verifying progress and performance					
Asking for feedback					
Accepting feedback					
Asking for ideas, help or solutions					
Informing about issues or mistakes					
Speaking up with ideas					

Acknowledging own mistake					
<b>Familiarity Behaviours</b>					
Talking about personal, non-work matters					
Making or laughing about a joke					
<b>Total Observed Behaviour</b>					
Categories indicating high psychological safety: (voice behaviours, supportive behaviours, learning or improvement behaviour and familiarity behaviours)					
Categories indicating lower psychological safety: (defensive voice behaviours, silence behaviours and unsupportive behaviours).					

---

## APPENDIX VII – OBSOLETE AND MISSING BEHAVIOURS

---

### OBSOLETE BEHAVIOURS

Thirdly, not all behaviours in the observation scheme seem necessary: As mentioned above, the behaviours *Talking about personal, non-work matters*, *Facial expression or body language indicates fear*, and *Asking a question (other)* were hardly used throughout the research. While *Talking about personal, non-work matters* has not appeared during the observed meetings is probably coincidental, there are legitimate reasons why *Facial expression or body language* and *Asking a question (other)* were not coded.

*Facial expression or body language indicates fear* is a behaviour that is very difficult to observe. It is very subjective when somebody is anxious and, therefore, the researcher was rather reluctant to use this code. This subjectivity also relates to the fact that it is a silent behaviour and these kinds of behaviour have already been found to be harder to observe reliably (Meinecke et al., 2016). It is also quite a bold statement to code someone as fearful just based on what one sees in a video. While it does make sense that this behaviour would be related to psychological safety, it might be excludable from the observation scheme due to the high subjectivity in coding this behaviour.

*Asking a question (other)* was added to the observation scheme by the researcher to account for questions that do not fit with the other *Asking ...* codes. However, the research shows that this code is hardly used, so most questions fit under other behaviours. Moreover, it might be questionable whether all questions need to be coded because they might not all relate to psychological safety. Therefore, this code might not be necessary and could be excluded from the observation scheme.

Moreover, since the observation scheme is very extensive, it is inevitable that there are some codes in the observation scheme that seem to overlap. These are *Providing information* and *Sharing procedures, knowledge and experience*, and *Providing positive feedback* and *Acknowledging achievements/congratulating*. Also, *Asking for feedback* could be similar to *Asking for ideas, help or solutions*, particularly *Asking for help*. There are small differences between all of these pairs but it is questionable whether these nuances are valuable to psychological safety research. It could be suggested to make *Sharing procedures, knowledge and experience* a part of *Providing information*, *Acknowledging achievements/congratulating* a part of *Providing positive feedback* and *Asking for feedback* a part of *Asking for ideas, help or solutions*. This would already reduce the complexity of the observation scheme.

---

### MISSING BEHAVIOURS

During the naked-eye observations, there was one team with low survey-measured psychological safety but the observational quantitative results were indicative of high psychological safety. In this team, qualitatively, a lot of gossiping was observed which could explain the low psychological safety score. Gossiping could lead to people being afraid that

their input will also be talked about behind their back thus decreasing psychological safety. Such behaviour is not yet included in the observation scheme. This might be a valuable addition and would allow the quantitative observations to also capture such behaviour.

Furthermore, there was a particular type of behaviour for which no code could be found during the computer-aided observations: a sort of negligence or indifference; Utterances such as “I don’t care”, “It doesn’t matter to me” or “I don’t know” that could have been interpreted as the person not wanting to put in the effort to take part in considering ideas or resolving issues. This was now coded as *Evading confrontation* which is a form of *Defensive Voice*, following the conceptualizations of Van Dyne et al. (2003). However, the behaviour of negligence and indifference would actually fall under the category of *Acquiescent Voice*. During coding, both of the researchers agreed that these indifferent utterances can give the speaker the impression that their input is not valuable enough, thus potentially decreasing psychological safety. Adding this to the observation scheme could give more insights into the nuances of psychological safety, as this type of behaviour is inherently different from *Defensive Voice* but still relates negatively to psychological safety.