Computers in Human Behavior Reports 14 (2024) 100413

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Psychological safety in interdisciplinary virtual student project teams: A validation study

Eline Rødsjø [a,*], Ela Sjølie [a], Peter Van Petegem [b]

[a] Norwegian University of Science and Technology, NTNU Faculty of Economics and Management Sem Sælands veg 1, 7034, Trondheim Norway
[b] University of Antwerp Faculty of Social Sciences Prinsstraat 13, 2000 Antwerpen, Belgium

ARTICLE INFO

Keywords:
Psychological safety
Validation
Multilevel confirmatory factor analysis
Virtual student teams

ABSTRACT

Psychological safety is an important variable for understanding teams, as it is positively associated with team performance, team learning, and creativity. A large body of research has been conducted on psychological safety in teams that work face-to-face. The aim of this study was to investigate the construct validity of Edmondson's Team Psychological Safety Scale in the context of virtual student teams working on a joint project. The sample used in this study consisted of 344 students who took part in an interdisciplinary project-based master's course at a Norwegian university. The data were analyzed with single-level confirmatory factor analysis (CFA) and multilevel CFA (MCFA). Reliability was investigated with intra-class correlation coefficients, in addition to Cronbach's alpha and McDonald's omega. Overall, the reliability estimates and goodness-of-fit estimates from the CFAs demonstrated an adequately good level of fit for the one-factor structure of psychological safety in the context of virtual student project teams. The Team Psychological Safety Scale was shown to have acceptable levels of construct validity for the virtual student project teams.

1. Introduction

Organizational researchers have found that psychological safety is an important component when trying to understand what makes teams work well together (Edmondson, 1999, 2004). Psychological safety can be described as shared trust among the team members that they will not punish, reject, or embarrass other team members for expressing themselves freely or speaking up in the group (Edmondson, 1999). Much research has been conducted on psychological safety in organizations, and past studies have indicated that working in a team environment that people perceive to be psychologically safe makes the team members more inclined to take interpersonal risks. Such safety is also positively associated with team performance (Baer & Frese, 2003; Edmondson, 1999; Fransen et al., 2011; Frazier et al., 2017) and has been shown to facilitate positive outcomes of team conflict on performance among teams with a high level of psychological safety (Bradley et al., 2012). In addition to the focus on psychological safety and team performance, researchers have also studied the effects of such safety on team learning. High levels of psychological safety are positively associated with learning behavior (Edmondson, 1999) and can facilitate creativity, enable team members to speak up, and create a climate where team members provide feedback to each other (Edmondson, 2004).

In part accelerated by the COVID-19 pandemic, virtual teamwork has become more widespread over the last few years. Most teams use digital tools to some extent, but virtual teams differ from face-to-face teams in that the members of the team only see each other on the screen and communicate exclusively through digital tools. As a result, virtual teams do not necessarily have access to all the same social and emotional cues in body language as teams that meet face-to-face (Edmondson & Daley, 2020; Marra et al., 2020; Moffett et al., 2023). Research shows that virtual collaboration brings challenges related to the social dimensions of teamwork, through impeded social interaction (Janssen & Kirschner, 2020; Sjølie et al., 2022), more formal communication (Pérez-Mateo & Guitert, 2012), and more misunderstandings within the team (Usher & Barak, 2020). Studies have also shown that some aspects of remote work hinder teams' psychological safety (Fleischmann, 2023; Lee, 2021; Tkalich et al., 2022). As a consequence, the psychological safety construct might work differently in a virtual context (McLeod & Gupta, 2023), and by extension also in virtual student teams.

While the body of research on psychological safety in virtual teams is growing, there is still a need for more studies on this important topic as the virtual mode of collaboration continues to gain momentum. The

* Corresponding author.
E-mail addresses: eline.rodsjo@ntnu.no (E. Rødsjø), ela.sjolie@ntnu.no (E. Sjølie), peter.vanpetegem@uantwerpen.be (P. Van Petegem).

https://doi.org/10.1016/j.chbr.2024.100413
Received 28 September 2023; Received in revised form 26 March 2024; Accepted 8 April 2024
Available online 9 April 2024

development towards more virtual teamwork also holds through for education, with increased use of team-based approaches in online learning environments (Farnell et al., 2021; McLeod & Gupta, 2023; Sjølie & van Petegem, 2022). There is thus a need for more research to properly understand how psychological safety develops in student teams who collaborate virtually. An early study by Zhang et al. (2010) on online discussion forums at a university in Hong Kong showed that high levels of psychological safety were positively associated with the intention to continue sharing knowledge among virtual student teams. In another study on engineering project teams, Gibson and Gibbs (2006) found that innovation in a sample of was negatively affected by working virtually but that this effect could be moderated by psychological safety. A more recent study, by Glikson and Erez (2020), looks into psychological safety in virtual student project teams and finds that the first relationally-oriented message sent in the team has a positive impact on the psychologically safe communication climate, while a study on student engineering teams by Cole et al. (2022) indicates that psychological safety might take longer to establish in a virtual setting. In other words, psychological safety has a considerable impact on teamwork, regardless of whether the teams are working face-to-face or virtually.

Based on the current trend towards widespread virtual teamwork and on extant knowledge that psychological safety is an important variable for understanding teams, this study aims to add to the existing literature on psychological safety by validating the construct in this new context of virtual student project teams. The construct validity of psychological safety has been studied intensively in face-to-face work teams (Newman et al., 2017), but how the construct works in virtual student teams remains unknown. Validation is an important part of research, as it lets researchers investigate the extent to which the measures used in the study measure the intended phenomenon. Validation is also central in the process of evaluating whether study results are reliable and generalizable to the population outside the sample included in a specific study. Construct validation should be an ongoing process, and best practice is to validate instruments whenever constructs are used in new populations and contexts. The current practice in some research areas, for instance social psychology (Flake et al., 2017; Hussey & Hughes, 2020), does not appear to be following what is generally considered best practice with validation. Flake et al. (2017) notes that there is a lack of ongoing validation when exiting scales are used in new contexts or populations, and that the practice of reporting evidence of construct validation could be increased. Validating the Team Psychological Safety Scale in the context of virtual student teams is an essential step to enable research on psychological safety in virtual student teams. Once the construct is validated, researchers will then be able to look into how psychological safety in virtual student project teams plays into a wide array of other group phenomena such as team performance, creativity, and learning, to name a few. The results also hold implications for further studies on psychological safety in virtual teams more broadly. The central research question of this study is, "What is the construct validity evidence for the Team Psychological Safety Scale in the context of virtual interdisciplinary student project teams?"

1.1. Psychological safety

Psychological safety is rooted in organizational research from the 1960s, where it was generally discussed as a prerequisite for feeling secure and capable of adapting one's behavior in the workplace when faced with organizational changes (Schein & Bennis, 1965). Researchers later viewed psychological safety in the context of learning as a tool to help members of a team overcome their defensiveness or learning anxiety in order to focus on problem-solving and the overall goal of the work (Edmondson & Lei, 2014).

Whereas the earlier work by Schein and Bennis focused on psychological safety as the individual perceives it (Schein, 1993; Schein & Bennis, 1965), Edmondson later defined psychological safety as a shared belief among team members that the team is safe in taking interpersonal risk (Edmondson, 1999). If a team experiences high degrees of psychological safety, then the members are more likely to feel that they can bring up issues with their team, ask others for help, and not be afraid that the team will use their mistakes against them (Edmondson, 1999). Researchers continue to examine psychological safety at both the individual and team levels, as well as within organizations. Edmondson's work (1999, 2004) marked the beginning of the research field of psychological safety at the team level (Edmondson & Lei, 2014), and many researchers have since used her seven-item Team Psychological Safety Scale to study psychological safety. A meta-analytic review by Frazier et al. (2017) concluded that future research should focus more on the team-level. The analyses in this study will follow Edmondson's line of team-level research on psychological safety.

According to Edmondson (1999), psychological safety facilitates team learning through behaviors such as sharing information, asking for feedback, discussing differences, and posing questions. Psychological safety has also been found to be a predictor of creative team performance (Kessel et al., 2012), mediated by the sharing of information and know-how. Teams that experience high levels of psychological safety are more willing to contribute to the team with ideas and actions (Edmondson & Lei, 2014), which might help explain why psychological safety is widely seen as facilitating increased knowledge sharing and positively influencing a team's creativity and innovativeness (Newman et al., 2017).

Researchers have also explored the relationship between psychological safety and the process of sharing ideas in teams among student teams. A study on engineering design student teams found that psychological safety was positively related to the quality of the ideas a team developed but negatively related to the number of ideas they developed (Cole et al., 2021), indicating that psychological safety might not increase efficiency by the quantity of ideas produced but rather the quality, thus making the team able to move forward in the process. Another study looked into the trajectory of psychological safety in engineering student teams to identify potential factors that affect both the development of and decreases in psychological safety (Miller et al., 2019, August 18-21). The teams in the study showed large variations in the trajectories of their psychological safety, but concept generation and the selection of ideas seemed to be critical points in the development of psychological safety. The qualitative data from the same study suggested that collaboration, respect for others' ideas, and anticipating variations in opinions could contribute to increased psychological safety, while lack of communication, deficiency of focus, disrespect for others' ideas, and interpersonal tension were linked to a decline in psychological safety (Miller et al., 2019, August 18-21).

The focus of the present study is on virtual student teams. Such teams refer to groups of students who interact through various communication technologies to accomplish their common goals (Johnson et al., 2002). Virtual environments tend to be designed with productivity in mind, with less emphasis on fostering social interactions (Abedin et al., 2011; Balacheff et al., 2009), while social interaction is known to be essential for collaborative learning (Johnson & Johnson, 2009) and important for fostering psychological safety. The number of studies on psychological safety in virtuals tudent teams has grown since the covid pandemic, but there is still not a large body of research with this focus. Gibson and Gibbs's (2006) early study on virtual teams showed a negative effect of working virtually on team innovation, but also noted that this could be moderated by psychological safety. A recent study by Moffett et al. (2023) on medical students working on an online design thinking project also finds the online environment to pose some barriers for teamwork, but as Gibson and Gibbs (2006), they conclude that psychological safety can mitigate some of this by putting in enough time and effort to nurture psychological safety. They suggest the use of break-out rooms and setting guidelines for the teamwork as promising tools to foster psychological safety. Interventions such as icebreakers and providing opportunities to engage in informal social interactions is supported by findings from both Moffett et al. (2023) and Cole et al.

(2022). Additionally, team performance in virtual teams seems to be positively affected by psychological safety, especially in teams with an abundance of national diversity (Kirkman et al., 2013). Zhang et al. (2010) showed that high levels of psychological safety had a positive relationship with team members' intention to continue sharing knowledge in virtual student groups, and research by Fleischmann et al. (2023) finds continued support for the relationship between antecedents such as interpersonal relationships, team dynamics and peer support on psychological safety in their sample of virtual student teams.

1.2. Validity

In general, team psychological safety is measured at the individual level, but the construct itself resides at the team level (Fig. 1). The individuals in the team answer the questions presented in the scale, but their answers point to their perception of the team as a whole and the climate that exists within the team. Psychological safety thus is an emergent multilevel construct, one where a higher-level phenomenon emerges from interactions that take place at a lower level (Jebb et al., 2017).

For a measure to accurately capture a construct, the construct should behave similarly at the individual and team levels, although the interpretation of the construct can differ according to the level of measurement and the level of aggregation (Forer & Zumbo, 2011). When more than one level of analysis is at play, multilevel validation techniques are necessary (Forer & Zumbo, 2011; Jebb et al., 2017).

The validation process in this study follows the multilevel validation framework presented by Chen et al. (2005), with a few additional approaches, as summarized by Jebb et al. (2017). Chen et al. (2005) distinguished five steps in their framework: construct definition, articulation of the nature of the aggregate construct, psychometric properties of the construct across levels of analysis, construct variability between units, and construct function across levels of analysis. The additional approaches for multilevel construct validation, as described by Jebb et al. (2017), are score similarity (reliability and agreement), psychometric isomorphism, and nomological network. In the following sections, we will briefly describe the different steps and apply them for the validation of Team Psychological Safety.

Step 1: Construct definition. The first step includes focusing on what the essence of the construct is, notably what is included and what is excluded. The principles of construct definition are the same for measures at the individual-level and multilevel constructs (Jebb et al., 2017). Within construct definition, researchers examine the theoretical underpinnings of the measure, conduct literature reviews, and interpret the concept to identify the depth and breadth of the construct they aim to study. For quantitative researchers looking to develop a measurement instrument, this phase also includes item development (Loevinger, 1957).

Researchers have already defined and differentiated psychological safety from similar concepts, such as group cohesion and trust (Bradley et al., 2012; Edmondson, 2004). Edmondson (1999) developed the Team Psychological Safety Scale in the context of work teams in a manufacturing company. As is generally considered good practice, later research tested the validity of the measure in other contexts as well. In their systematic review, for example, Newman et al. (2017) found the measure to be reliable across diverse samples; researchers have also validated the measure among different samples, including Brazilian workers (Ramalho & Porto, 2021), student engineering design teams (Miller et al., 2019, August 18-21), and health-care teams (Kessel et al., 2012).

Step 2: Articulation of the nature of the aggregate construct. The second step is to articulate the nature of the aggregate construct and specify a composition model for the construct. A composition model specifies the functional relationship within a phenomenon or construct at the different levels of analysis and is determined by the theoretical nature of the construct (Jebb et al., 2017). A phenomenon can be aggregated, or it can emerge from lower-level interactions (e.g., at the individual level), in different ways to become a higher-level phenomenon, for example at the group level (Jebb et al., 2017). Different composition models hold different assumptions about how this process takes place.

Chan (1998) describes five different composition models: additive models, direct consensus models, reference-shift consensus models, dispersion models, and process models. In an additive model, the higher-level unit is a summation of the lower-level units. In this type of composition model, consensus among the lower-level units is not required (Chan, 1998). In a direct consensus model, the meaning of the higher-level construct is found in the consensus among the lower-level units, and the measurement items reference the lower level only (Chan, 1998; Jebb et al., 2017). For the reference-shift composition model, researchers assume that the individuals in the group can perceive a shared "global property" (Jebb et al., 2017). In this type of composition model, researchers not only are interested in individuals' own perception of the construct, but also in how they believe others in the group perceive the construct (Chan, 1998). In dispersion models, the construct emerges as the variability of within-group units and may be operationalized as either the group standard deviation (SD) or as the variance of range (Chan, 1998). Process models are often used to describe the process in which an organization or unit moves from lack of within-group agreement to a state of high within-group agreement (Chan, 1998; Jebb et al., 2017).

Psychological safety is a team-level construct that emerges from individuals' own perception of the team, as well as their beliefs about how the other team members view the level of psychological safety. One could say that a certain level of psychological safety is present in the team that the individuals in the team can feel and perceive. The individual scores are aggregated to become a team score by averaging the responses for all the individuals in the team for each item in the

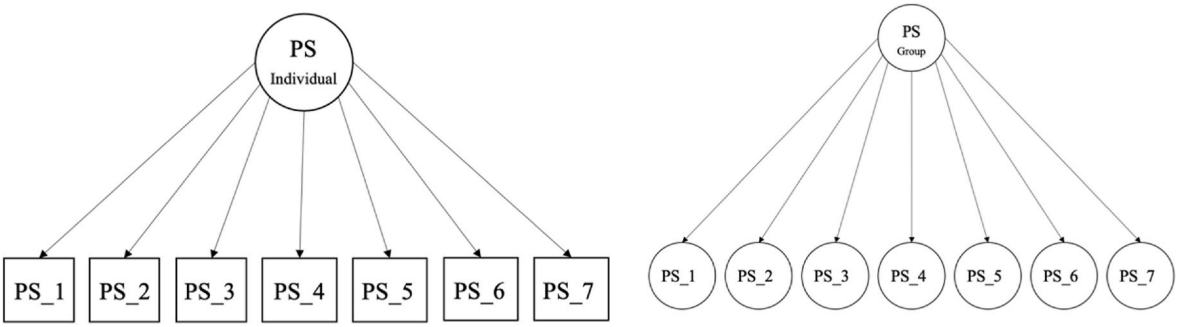

Fig. 1. Individual-level model (measurement level) and team-level model (construct level) of the Team Psychological Safety Scale.

psychological safety measure. This approach means that the composition model that best describes the relationships among the lower and higher levels of analysis of psychological safety is a reference-shift composition model. See Chan [1998] and Jebb et al. [2017] for a more in-depth review of the different composition models.

Step 3: Psychometric properties of the construct across levels of analysis. The third step in Chen et al.'s (2005) framework is to test the construct's psychometric properties across levels of analysis. The construct's composition models determine what type of empirical validation strategy is required, and referent-shift models require that researchers look into all the additional approaches of score similarity, psychometric isomorphism, and nomological network both with and without homology (as described by Jebb et al. [2017]), in order to meet the criteria for construct validity.

Score reliability. Two kinds of reliability should be investigated for multilevel analysis. The first is psychometric reliability, the classic reliability that expresses the internal consistency and repeatability of the scores. This type is tested with Cronbach's alpha and McDonald's omega. Alpha is computed with the inter-item covariance within a scale, the variance of the scale score, and the number of items included in the scale (Geldhof et al., 2014). McDonald's omega represents the ratio of a scale's estimated true score variance, relative to the scale's total variance, and is in that way conceptually similar to Cronbach's alpha. In contrast to Cronbach's alpha, the omega coefficient also acknowledges the possibility that the items in the scale could have heterogenous item-construct relations, whereas Cronbach's alpha which assumes that all items have the same factor loading, i.e., the same importance to estimate the construct. The conventional cut-off level of both Cronbach's alpha and McDonald's omega is 0.7 (Geldhof et al., 2014).

The second approach to multilevel reliability analysis is aggregate reliability. This step is important when dealing with a referent-shift composition model, as the scores within the groups must be similar enough to justify aggregating the scores from the individual level to say something about the group level (Jebb et al., 2017). Aggregate reliability is tested with inter-rater agreement (IRA) and inter-rater reliability (IRR). IRA is used to demonstrate the similarity of within-group responses to check if the assumption of the reference-shift composition model is justified, namely that the data from the individuals in each group are similar enough to justify aggregating the data to the group level (Jebb et al., 2017). IRR is used in multilevel measurement to provide an estimate of the reliability of the higher-level construct, and to demonstrate the degree of similarity within groups to justify aggregation (Jebb et al., 2017). Inter-rater agreement and reliability can be estimated with intra-class correlation coefficients (ICCs), called $\text{ICC}_{(1)}$ and $\text{ICC}_{(2)}$. These coefficients are well suited for multilevel analysis because the tests presume that the lower-level units (e.g., individual scores) are nested within higher-level units, such as groups (Jebb et al., 2017). $\text{ICC}_{(1)}$ can be interpreted as the proportion of the total variance that can be attributed to the fact that the lower-level units are nested within higher-level units (in this case, the individual scores and the student teams, respectively). $\text{ICC}_{(2)}$, in contrast, is a reliability estimate used to show that the group means can be seen as reliable indicators of the higher-level construct (Jebb et al., 2017), which in the present study means the teams' overall level of psychological safety.

Factorial validity and psychometric isomorphism. After establishing the construct's reliability, examining the factorial validity is the next step. Looking at the factor structure is an important part of multilevel construct validation when higher-level measures are made up of multiple items (Jebb et al., 2017). The factor structure describes the relationship between the items used to measure the construct and the latent variable(s) and is the most important aspect of factorial validity.

Two types of factor analysis may be distinguished: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). While EFA is commonly used to discover the factor structure of a measure, CFA is used to assess a defined measurement model of a construct (Brown & Moore, 2012) by testing the hypothesis that the items (such as responses to survey questions) are associated with specific factors (Byrne, 2011).

As opposed to the conventional single-level analysis of factor structure, psychometric isomorphism is the best approach when testing the factorial validity of a multilevel construct. Researchers use psychometric isomorphism when they need to demonstrate that a factor structure is consistent across levels (Jebb et al., 2017). With a reference-shift composition model, such as with psychological safety, the meaning of the scores at the lower level must be sufficiently maintained at the higher level. Researchers can check if this is the case by looking into whether the factor structure is preserved across levels (Jebb et al., 2017).

Psychometric isomorphism may be tested in various ways, using either separate estimation or simultaneous estimation (Jebb et al., 2017). Using separate estimation, researchers estimate the factor structures at the different levels separately, then make one independent model with the raw individual data and another with the aggregated group-level data (Jebb et al., 2017). Separate estimation is generally not recommended, due to concerns that doing so can lead to biased parameter estimates and incorrect standard errors (see Muthén, 1994). Simultaneous estimation, in contrast, takes the nested structure of the data into account.

Confirmatory factor analysis. In this study, CFA is used to analyze whether the data collected among virtual student project teams fit the predefined factor structure of the Team Psychological Safety Scale on both the individual and team levels. The fit of a CFA model can be assessed at different levels: the overall model, the equation, and the parameter level (factor loadings). The suggested model fit criteria are as follows: $\chi 2$, comparative fit index (CFI) and Tucker–Lewis index (TLI) of 0.90 or above, root mean square error of approximation (RMSEA) between 0.05 and 0.10, and a standardized root mean square residual (SRMR) value of 0.08 or lower (Bentler, 1990; Byrne, 2011; Hu & Bentler, 1999). The x^2 fit criteria is used to compare the observed sample data with the expected data and check whether the difference is statically significant, while the CFI is used to examine the discrepancy between the sample data and the hypothesized model (Hu & Bentler, 1999). CFI also adjusts for issues relating to the sample size, which the x^2 does not. TLI is a relative fit index that is used to analyze the discrepancy between the values of $x^2$of the hypothesized model and the null model, and the RMSEA test calculates the fit of the model compared to the population's covariance matrix (Byrne, 2013). SRMR calculates the standardized difference between observed and predicted correlation (Hu & Bentler, 1999).

Multilevel confirmatory factor analysis. Multilevel confirmatory factor analysis (MCFA) is an extension of CFA that can be used to analyze the individual level and the group level simultaneously. Kyriazos (2019) summarizes three steps necessary to incorporate the multilevel approach into MCFA.

The first step is to examine the ICCs of the items to check how much of the variance in the items is explained by the individuals' group membership: the higher the ICC score, the more of the score variance is attributed to the grouping variable. The ICC values of the observed variables calculated at this stage of the analysis should be above 0.10 in order to justify adding the nested structure of the data into the model.

The second step is to analyze the data of the model's lower level. A standard CFA is used for this step, and the fit of the model is determined with conventional goodness-of-fit criteria. If the fit of the CFA model of the data at the individual level is deemed to be satisfactory, then the analysis can proceed to the next step.

The third and final step is to test the factor structure of both the individual and group levels simultaneously (Kyriazos, 2019). The goodness-of-fit indices of this multilevel CFA show the extent to which the specified model fits both the within-group model data and the between-group model (Fig. 2). One factor to note is that the usual goodness-of-fit indices should be interpreted less strictly for the group

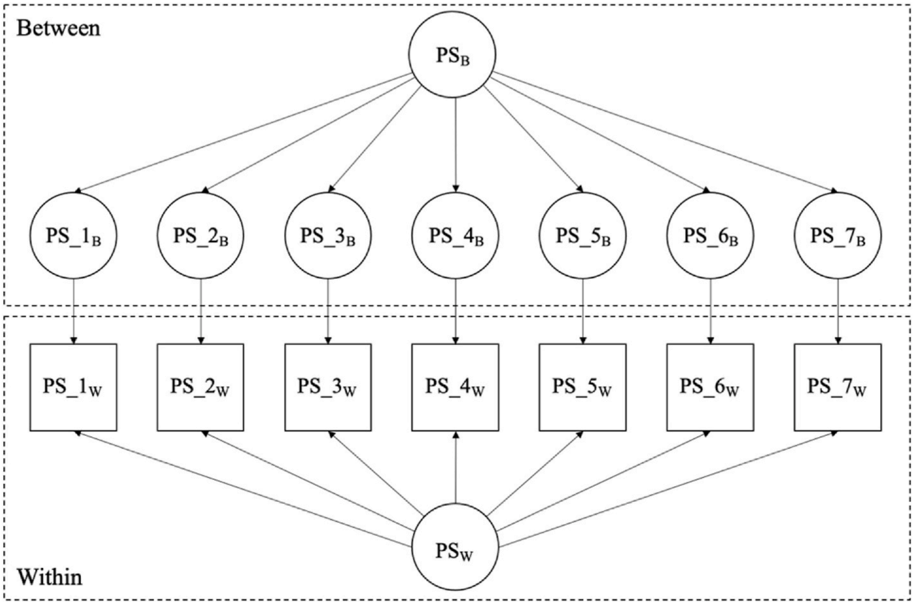

Fig. 2. Within-group and between-group levels of the Team Psychological Safety scale.
Note: Circles indicate latent variables, and squares indicate observable variables. The "within" level represents the variation within the individuals at the lower level of the multilevel model. The "between" level of the model represents the groups at the higher level of the multilevel model. The seven items in the scale are pictured as observable variables at the "within" level, as they represent the direct responses from the individuals. At the group level, the items are no longer directly observable, as they are a function of the answers from multiple group members.

level of the MCFA, as such indices are less sensitive to misspecifications at the group level of the model than they are at the individual level (Dyer et al., 2005; Maas & Hox, 2005).

Step 4: Construct variability between units. As previously mentioned, in order to aggregate the data to a higher level for constructs with a reference-shift or consensus model, it is necessary that the members within a group to some extent have a shared perception of the phenomenon being studied. Researchers should also investigate the variability between units, or between groups. To some extent, distinguishing the groups from one another should be possible if the construct captures something unique relevant to each group's perception of themselves.

Researchers do, however, continue to debate whether variability is strictly necessary if agreement exists within the units (Fischer, 2015). If all individuals in all the groups included in the sample experience high levels of psychological safety, then agreement will exist within units, but the variability will be scant between the units. With no variability between groups, identifying group effects in multilevel models becomes difficult (Fischer, 2015). If the groups under consideration all come from the same organization, for example, then the lack of variability could be caused by organization-level effects.

Step 5: Construct function across levels of analysis. In this step of the validation process, the construct is investigated in light of similar and differing constructs and their associated antecedents and outcomes. Cronbach and Meehl (1955) introduced the concept of nomological networks as a representation of the theoretical construct (or constructs) and framework of a study, the empirical framework for how it is to be measured, and the relationship between the construct and the empirically observable manifestations of the construct and other constructs. Nomological validity evidence describes the extent to which a specific construct behaves as hypothesized within a system of related constructs (Jebb et al., 2017). The latter step is beyond the scope of this study and thus will not be discussed in this paper.

2. Methods

2.1. Context

The data of this study were drawn from fourth-year university students who participated in a master's course called "Experts in Teamwork" at the Norwegian University of Science and Technology (NTNU). In this project-based course, students are organized into classes of about 30 students in each, and then into interdisciplinary teams of four to six students. The students come from different faculties, including information technology, engineering, natural sciences, health sciences, economics, social sciences and architecture and design. The focus of the course is for the students to develop interdisciplinary teamwork skills while working together on a self-defined, real-world project throughout the semester. No specific guidelines are provided regarding how to distribute team roles and tasks. While all classes focus on interdisciplinary teamwork skills, each class has its own overarching project topic. Some examples of such class-topics include "Creating value from waste", "Digital communities and welfare", "Smart energy management" and "Ocean justice". Based on the class-topic, the student groups chose themselves the specific question or problem they want to address in their project. The students can choose to enroll in a class with on-campus collaboration or participate in a class where all teamwork takes place virtually. These virtual classes have students from three different campus cities in Norway. The students can also choose whether to sign up for a semester-based version, where they meet their team once a week for a whole semester, or an accelerated version of the course, where they meet every day for three consecutive weeks. The course is taught in both Norwegian and English and includes both Norwegian and non-Norwegian exchange students.

2.2. Research design

The Team Psychological Safety Scale was administered as part of a larger survey that involved additional measures. All students enrolled were invited to fill out a survey at three time points during the course: the beginning (T1), middle (T2), and end (T3). The psychological safety measure was included as a part of the survey at T1 and T3. This study's analysis is based on the data from T3. The first couple of days of the course are usually filled with information and lectures on the class-topic, and the students are not allocated much time in their teams before day 2 or 3. The survey at T1 is usually on day 2 or 3, so not all teams had been able to interact sufficiently with each other at T1 to measure psychological safety at this point.

The study was approved by the Data Protection Office for Research at the Norwegian Centre for Research Data, or NSD (#749025). The participants were asked to complete an informed consent form before participating in this study and could withdraw at any time. The data used in this study were fully anonymized prior to the analyses.

2.3. Sample

The sample for this study consisted of 344 (out of 380) students who took part in the virtual semester-based version of the course. The sample included data from 70 different student teams. The average number of participants in a student team was 5.37 (SD = 0.66). The gender distribution of the participants was 215 (59.6%) men, 126 (34.9%) women, and 6 (1.7%) other/preferred not to say; 14 (3.9%) subjects did not answer the question. The grade point average (GPA) of the students was 2.21 (SD = 0.70), which corresponds to a letter grade between C and B.

2.4. Measure

The Team Psychological Safety Scale used in this study was developed by Edmonson (1999) and consists of seven items. All items were answered on a five-point Likert scale (1: totally disagree, 2: disagree, 3: neither disagree nor agree, 4: agree, 5: totally agree). The items were translated from English to Norwegian by one translator and then back-translated to English by another translator. The back translation was then compared with the original items to evaluate whether they had retained their original meaning throughout the translation process. See Appendix A for the English and Norwegian items. Roughly two-thirds (64.3%) of the students answered the English version of the survey, while 35.7% answered the Norwegian version.

2.5. Data analysis

Data from the 344 students who participated in the virtual version of the course were extracted from the survey database and saved as a separate data file. The reversed survey items were recoded, and group averages for the 70 student teams were then calculated. From a theoretical stance, psychological safety is situated at the group-level. This makes averaging at the team level a suitable choice for aggregation. While weighted averaging could have been considered, the lack of criteria and data to assign weights to individual responses rendered it less feasible in our context. Similarly, while a consensus method within the group could have been explored, unfortunately, such data were not available. Additionally, a multilevel approach might have been viable given larger sample sizes at the lowest level; however, the small size of our student teams precluded us from obtaining sufficiently stable estimates through this method. Teams with only one respondent were removed from the dataset, which led to the removal of two teams, resulting in a total of 68 observations (student teams) in the dataset used for the analyses.

Descriptive statistics (mean, SD, skewness, and kurtosis) were calculated for each item in the measure, for both the individual and group levels (see Tables 1 and 2). The analysis was run using SPSS Statistics version 28.0.1.0 (142). Cronbach's alpha and McDonald's omega were calculated in SPSS for the single-level CFA to test the reliability of the psychological safety measure. For the multilevel CFA, McDonalds's ω was calculated manually in Excel using the MPlus output from the multilevel analysis. The calculations followed the equations presented in Geldhof et al. (2013). The intra-class correlation coefficients $ICC_{(1)}$ and $ICC_{(2)}$ were also calculated.

Table 1
Descriptive statistics for Team Psychological Safety items, individual level.

Item	N	Mean	SD	Skewness	Kurtosis
01	344	4.52	0.76	−1.848	3.772
02	344	4.09	0.81	−1.009	1.574
03	342	4.71	0.68	−2.858	9.034
04	343	4.21	0.80	−0.984	1.364
05	343	4.52	0.84	−2.272	5.752
06	344	4.36	1.10	−1.948	2.946
07	342	4.24	0.77	−1.133	2.135

Note: SD = Standard deviation.

Table 2
Descriptive statistics for Team Psychological Safety items, team level.

Item	N	Mean	SD	Skewness	Kurtosis
01	68	4.594	0.350	−1.227	1.647
02	68	4.100	0.457	−0.505	−0.095
03	68	4.765	0.307	−2.050	3.949
04	68	4.202	0.445	−0.279	−0.209
05	68	4.591	0.367	−1.564	4.109
06	68	4.345	0.591	−0.649	−0.718
07	68	4.249	0.428	−0.934	1.029

Note: SD = Standard deviation.

A single-level CFA was conducted, using MPlus 8, to test the structural validity of the measure at the lower level of the construct. Psychological safety was defined as the latent variable, and the seven scale items were added to the model as observed variables (Fig. 2), using robust maximum likelihood as the estimation method, given the non-normal distribution of the items. The suggested model fit criteria were as follows: CFI and TLI of 0.90 or above, RMSEA between 0.05 and 0.10, and SRMR values of 0.08 or lower (Bentler, 1990; Byrne, 2011; Hu & Bentler, 1999).

MCFA was then conducted to test the structural validity of the full model. In MCFA, the total covariance matrix is parted into two components (the "within" and "between" groups) in order to test whether the factor structure remains constant across both the individual and team levels. The multilevel model was specified so that the individual scores were nested according to the team variable (Fig. 2). Here, too, robust maximum likelihood was used as the estimation method.

Multiple goodness-of-fit indices were used to evaluate the models. The suggested model fit criteria were similar to those for the single-level CFA, acknowledging that the criteria should be interpreted in a less strict way at the group level of the MCFA (Dyer et al., 2005; Maas & Hox, 2005).

3. Results

3.1. Descriptive statistics

For the data on the individual level, the skewness statistics ranged from −2.86 to −0.98, while kurtosis values were between 1.36 and 9.04 (Table 1). The skewness statistics for the team level ranged from −2.050 to −0.505, and the kurtosis values were between −0.718 and 4.109 (Table 2). These findings indicate that the data are not normally distributed, and that the robust maximum likelihood estimation method should be used in the factor analysis.

3.2. Reliability

The $ICC_{(1)}$ values for the items of the psychological safety construct were between 0.074 and 0.149, with an average $ICC_{(1)}$ of 0.107 (see Table 3). The $ICC_{(1)}$ value for the mean psychological safety score of the participants was 0.182. The $ICC_{(2)}$ value was estimated to be 0.688. The

Table 3
Intra-class correlation coefficients (ICCs) of the items in the Team Psychological Safety Scale.

Item	ICC
01: If you make a mistake on this team, it is often held against you.	0.109
02: Members of this team are able to bring up problems and tough issues.	0.149
03: People on this team sometimes reject others for being different.	0.063
04: It is safe to take a risk on this team.	0.115
05: It is difficult to ask other members of this team for help.	0.074
06: No one on this team would deliberately act in a way that undermines my efforts.	0.099
07: Working with members of this team, my unique skills and talents are valued and utilized.	0.143

way $ICC_{(2)}$ is calculated in SPSS is not fully adaptable to multilevel data with varying numbers of “raters” (in this case, team members) for each case that has been rated (here: the team). Mean imputation was used to ensure that data from all 68 teams were included in the $ICC_{(2)}$ estimate.

For the single-level CFA, Cronbach’s α was 0.726 and McDonalds’s ω was 0.724, both values above the conventional level of 0.7. These values were calculated in SPSS. For the multilevel CFA, McDonald’s ω was 0.695 for the within-group level, and 0.936 for the between-group level (Table 4).

3.3. Single-level CFA model estimates

Review of the goodness-of-fit statistics (Table 4) showed a good fit for the single-level CFA. The chi-square value for the model was χ^2 (14) = 29.824, at $p = 0.008$. The CFI value was 0.938 and the TLI value was 0.907, indicating an acceptable level. The RMSEA value was 0.057, and the SRMR value was 0.039, both of which were below their respective recommended threshold values of RMSEA (<0.08) and SRMR (<0.10). The factor loadings for the items in the single-level CFA model were between 0.356 and 0.623, with an average standardized factor loading of 0.544, which indicates a moderately good fit (Table 5).

3.4. Multilevel CFA model estimates

The results of the MCFA for the chi-square values were as follows: χ^2 (28) = 53.322, at $p = 0.002$ (Table 4), CFI = 0.930 and TLI = 0.895, RMSEA = 0.051 (<0.08). The SRMR value for the “within” part of the multilevel model was 0.044 and for the “between” part was 0.147. The standardized factor loadings for the items in the “within” part of the model were between 0.310 and 0.681, with an average factor loading of 0.516. The standardized factor loadings of the between-group level were between 0.354 and 1.006, with an average factor loading of 0.768, indicating a good fit (Table 5).

4. Discussion

The aim of this study was to investigate the construct validity of the Team Psychological Safety Scale in the context of virtual student project teams. We did this by calculating reliability estimates and performing single-level and multilevel CFA. Goodness-of-fit statistics were obtained for both models. The results demonstrated that the models had adequate reliability and goodness of fit, which suggests that the Team Psychological Safety Scale can be used to measure psychological safety in virtual student project teams.

Table 4
Single-level and multilevel CFA model fit indexes.

	Omega	χ^2	*df*	CFI	TLI	RMSEA	SRMR
Single-level CFA	0.724	29.824	14	0.938	0.907	0.057	0.039
Multilevel CFA	W = 0.695 B = 0.936	53.322	28	0.930	0.895	0.051	W = 0.044 B = 0.147

Note: χ^2 = chi squared; *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; W = within; B = between; χ^2 was statistically significant at $p < 0.05$ for both levels.

Table 5
Factor loadings of the items in the Team Psychological Safety Scale.

Item	Single-level CFA	Multilevel CFA	
		Within	Between
01: If you make a mistake on this team, it is often held against you.	0.487	0.505	0.544
02: Members of this team are able to bring up problems and tough issues.	0.610	0.546	0.870
03: People on this team sometimes reject others for being different.	0.597	0.681	0.354
04: It is safe to take a risk on this team.	0.623	0.562	0.955
05: It is difficult to ask other members of this team for help.	0.518	0.480	0.948
06: No one on this team would deliberately act in a way that undermines my efforts.	0.356	0.310	0.698
07: Working with members of this team, my unique skills and talents are valued and utilized.	0.618	0.525	1.006

The aggregate reliability of the team level of psychological safety was assessed by calculating intra-class correlation coefficients (ICCs), which are used to estimate the proportion of the variance that can be attributed to the hierarchical structure of the data. The $ICC_{(1)}$ is used to asses the portion of the total variance that can be attributed to the nesting of lower-level units (i.e. individual scores) within higher level units (i.e. team scores), and should be above the threshold value of 0.10 (Kyriazos, 2019).

The $ICC_{(1)}$ values of the seven items used to measure Team Psychological Safety were on average above the suggested level of 0.10. Two of the items had $ICC_{(1)}$ values below 0.10 (item 03 = 0.063, and item 05 = 0.074). These $ICC_{(1)}$ values were below the suggested threshold value of 0.10, but as Hox (2013) notes, values between 0.05 and 0.10 can also indicate a clustering effect in the data, thus warranting the use of multilevel analysis. The $ICC_{(1)}$ value for the mean psychological safety score at the individual level was 0.182, well above the threshold value of 0.10. These reliability estimates indicate that a substantial portion of the overall variance in the scores may be attributed to the team level in the sample. This finding further means that the data showed sufficient between-group variation to be able to statistically warrant the use of a multilevel analysis of psychological safety in virtual student project teams.

The $ICC_{(2)}$ value, a reliability estimate used to check if the group mean can be considered a reliable indicator of the higher-level construct (Psychological Safety) was estimated to be 0.688, which is above the threshold level of 0.40 (Fleiss, 1986). This level indicates sufficient reliability of the group means.

4.1. Single-level CFA estimates

The psychometric reliability estimates for the single-level CFA were both above the suggested threshold level of 0.700, with Cronbach’s α = 0.726 and McDonalds’s ω = 0.724. The chi-square of the model was χ^2 (14) = 29.824, $p = 0.008$. The goodness-of-fit indices (CFI, TLI, RMSEA, and SRMR) all had values indicating that the model had an acceptable level of fit (Table 4).

All items but two had factor loadings above 0.5. Neither item 01, “If you make a mistake on this team, it is often held against you” (0.487), or item 06, “No one on this team would deliberately act in a way that undermines my efforts” (0.356), explained as much of the variance of the latent variable of psychological safety as the other items in the model. Some have argued that items with a factor loading of less than 0.4 should be dropped (Guadagnoli & Velicer, 1988). But with a relatively low sample size and a strong theoretical basis for keeping the lower-scoring item, the lower score of item 06 is not sufficient to recommend deletion. Overall, the average standardized factor loading was 0.544, which indicates that the factor structure of the model has a moderate to good fit.

Taken together, the reliability estimates, the goodness-of fit indices, and the factor loadings suggest that the single-level CFA model has adequate structural validity. The single-level CFA, however, is based on the total variance and does not take the hierarchical structure of the data into account. While these results give an indication that continued analysis of team psychological safety in virtual student project teams with a multilevel CFA is worthwhile, the single-level CFA estimates cannot on their own be used to evaluate the construct validity of the measure.

4.2. Multilevel CFA estimates

For the multilevel model, McDonald's omega was just below 0.700 for the within-group level, and well above this threshold number (0.936) for the between level. These reliability estimates indicate that the reliability is higher for the between-group level of the model than it is for the within-group level. The fact that the reliability estimates are higher for the between-group level is in line with the theoretical model of Team Psychological Safety as a team-level construct.

The estimates of the goodness-of-fit indices CFI and RMSEA indicated that the MCFA model had an acceptable level of fit (Table 4). The multilevel CFA model showed a value for χ^2 (28) = 53.322, at $p = 0.002$. The TLI value was slightly below 0.90 (0.895), which could indicate room for improvement in the model. The SRMR value was within satisfactory levels in the "within" part of the model at 0.440, but the value for the "between" part of the model was 0.147, which is above the recommended threshold of 0.08. Goodness-of-fit indices, however, should not be interpreted as strictly for multilevel models as for single-level models. These indices are more sensitive to misspecification of multilevel models, and a common practice for cutoff levels for multilevel CFA models has yet to be established (Dyer et al., 2005; Kyriazos, 2019; Maas & Hox, 2005). Because the TLI value was just below the suggested threshold, it should be considered adequate for the multilevel model.

The factor loadings of the multilevel CFA model indicated a moderate to good fit at the "within" level, with an average standardized factor loading of 0.516, and a good fit at the "between" level, with an average standardized factor loading of 0.768. At each respective level, one of the items had a lower factor loading. For the "within" level this was item 06 (0.310), "No one on this team would deliberately act in a way that undermines my efforts," the same item that showed the lowest factor loading in the single-level CFA model. At the "between" level, item 03 explained the least amount of variance (0.354): "People on this team sometimes reject others for being different." This finding is in line with the observed covariation at the team level, since item 03 had the lowest ICC value (0.063) of all the items. One possibility is that item 03 was less strongly connected to team-level perception of psychological safety than the individual level. The question could also effectively be measuring something other than psychological safety when applied to a team level for virtual student teams.

4.3. Comparison of single-level CFA and multilevel CFA estimates

Both the single-level CFA and the MCFA model showed adequate levels of psychometric reliability. The McDonald's omega value was lowest for the within-group level of the MCFA and highest for the between-group level. The single-level CFA model had a high enough McDonald's omega value to indicate sufficient reliability, but it was lower than the between-group level of the MCFA. This finding supports the notion that psychological safety was a team construct in the virtual student project team sample used in this study.

The goodness-of-fit indices indicate a reasonable fit for both the single-level and the multilevel model. Although the goodness-of-fit indices were slightly lower for the multilevel CFA model than the single-level CFA model, the factor loadings had larger values in the "between" level of the multilevel model when compared to both the MCFA "within" level and the single-level CFA. Standardized factor loadings were consistently higher in the between-group level for six out of seven items, with an average factor loading of 0.768, as compared to both the within-groups level of the MCFA, with an average of 0.516, and the single-level CFA, with an average of 0.544. The higher factor loadings at the between-group level indicate that the multilevel model of psychological safety had a superior fit at this level in comparison with the within-group level. These results lend support to the use of multilevel analysis, and that Team Psychological Safety is a team-level construct.

A comparison of the single-level CFA and the "within" level of the multilevel model showed that the single-level CFA factor loadings were higher than those of the "within" level of the multilevel CFA. This finding likely resulted because the single-level CFA does not factor into the nesting of the data, as is the case for MCFA. The "group effect" of psychological safety is still present and affects the item responses at the individual level in the single-level CFA. To determine what the model fit indices would be for a model that contained only the variance of the individuals, and not the variance stemming from belonging to a team, one possibility would be to perform a separate factor analysis on the sample within-group covariance matrix. Based on this result, the goodness of fit of the factor structure could then be separately estimated at the within-group level. The same could be done with the between-group covariance matrix prior to running the multilevel CFA. The model fit indices from these analyses could then be compared to the MCFA estimates.

The difference between a single-level CFA and a factor analysis based on the sample within-group covariance matrix is that for the latter, the data are adjusted to remove between-group differences, which may be done by subtracting the relevant team means from the individual scores (Dyer et al., 2005). If the construct-relevant variance predominantly lies at the between-group level, then the model estimated using the within-group covariance would show a worse fit than the single-level CFA based on the total variance. These extra analyses will provide additional information about whether a team-level factor structure would be appropriate for the data. This approach is different from separate estimation—the method that Jebb et al. (2017) advised against—as the data are not only investigated separately at the two levels without accounting for the hierarchical structure, but each level is also investigated based on the sample within-group and between-group covariance matrix prior to running a simultaneous model.

4.4. Comparison of the results to those of other multilevel studies

Although comparing the goodness of fit of the models in this study with other validation studies of psychological safety could be useful, most researchers have either conducted the analysis in a different manner or adjusted the Team Psychological Safety Scale in some way. For example, Ramalho and Porto (2021) validated a psychological safety measure in the context of female Brazilian workers (N = 8310). They used an adapted version of Edmondson's Team Psychological Safety Scale with six items, translated into Portuguese. They first investigated the factor structure with principal component analysis and an exploratory factor analysis, and subsequently with CFA. The factor loadings in their one-factor single-level CFA model were higher than in this study, as they were all above 0.6 (ranging from 0.668 to 0.815), and their fit indices were better (CFI = 0.995, TLI = 0.992, RMSEA = 0.07, χ^2 (9) = 205,273). Their study left out the question of group-level analysis of the construct, however, whereas our findings indicate that the measure seems to work well.

To the best of our knowledge, no studies to date have applied MCFA with the Team Psychological Safety construct. Whitton and Fletcher (2014) conducted a multilevel CFA on the Group Environment Questionnaire, a measure of group cohesion. They concluded that the results provided support for the multilevel structure of the measure based on "low but acceptable fit statistics," with CFI estimates between 0.810 and 0.850 and RMSEA estimates between 0.080 and 0.110. These estimates are generally lower than those obtained in our analysis, indicating that when we used their benchmarks, the models in the present study had acceptable levels of fit as well.

4.5. Limitations and suggestions for further research

A few limitations should be considered when interpreting the findings of this study. The sample size was about 350 students, divided into 68 teams. The sample ideally should have been larger to enable us to run both an EFA and a CFA on half the sample. The virtual student project teams also came from different classes. A future analysis could account for this third level of clustering and investigate the extent to which teachers affect a teams' level of psychological safety.

In a next phase of the validation of the instrument it would be desirable to use a measurement invariance approach to verify to what extent the instrument behaves stable across groups, e.g., for English versus Norwegian version. Measurement invariance is "a property of a measurement instrument, implying that the instrument measures the same concept in the same way across various subgroups of respondents" (Davidov et al., 2014).

Another limitation is that the current study only used data from one time point, which was during the move to online work during the covid pandemic. One suggestion for further research could be to compare the development of psychological safety in pre- and post-pandemic samples to how new ways of doing virtual collaboration might affect this process.

Furter research should measure psychological safety at multiple time points and compare the results and look at different demographics. The students in the sample used in this paper worked on a broad project with minimal structure, and it would be interesting to see how the type of project can influence the development of psychological safety.

Another potential topic to examine is the development process of psychological safety levels in teams that have only ever worked virtually, and teams that have had an in-person "get to know each other" period prior to working virtually. More research on psychological safety in virtual student project teams could lead to expanded knowledge about what educators could do to help students foster psychologically safe teams. Furthermore, it would be interesting to examine differences between virtual student teams and team of working professionals when the Team Psychological Safety Scale is validated in the context of virtual student teams.

5. Conclusion

Psychological safety has become an important construct in research into how teams can work well together. Because virtual teamwork has become more common in both working life and education, it is necessary to examine the validity evidence of the construct in this new setting to enable further research on psychological safety in virtual teams. This study investigated the construct validity of the Team Psychological Safety Scale in the context of virtual student project teams.

The aggregate reliability estimates indicated that the aggregation of data from the lower level (individuals) to the higher level (teams) was justified. The goodness-of-fit indices also indicated an adequate model fit for both the single-level and multilevel confirmatory factor analysis, confirming that the one-factor structure of psychological safety also applies for the sample of virtual student teams. As for now, this result means that we can use the Team Psychological Safety Scale to continue to research psychological safety in virtual project teams.

CRediT authorship contribution statement

Eline Rødsjø: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Ela Sjølie:** Resources, Supervision, Writing – review & editing. **Peter Van Petegem:** Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. The Team Psychological Safety Scale (Edmondson, 1999)

Survey item	Survey items in English	Survey items in Norwegian
01	If you make a mistake on this team, it is often held against you. (*Reversed*)	Hvis du gjør en feil i dette teamet, blir det ofte brukt mot deg.
02	Members of this team are able to bring up problems and tough issues.	Medlemmene i dette teamet evner å ta opp problemer og vanskelige tema.
03	People on this team sometimes reject others for being different. (*Reversed*)	Personene i dette teamet avviser av og til andre fordi de er annerledes.
04	It is safe to take a risk on this team.	Det er trygt å ta sjanser i dette teamet.
05	It is difficult to ask other members of this team for help. (*Reversed*)	Det er vanskelig å spørre andre i teamet om hjelp.
06	No one on this team would deliberately act in a way that undermines my efforts.	Ingen i teamet ville med hensikt undergrave mine bidrag.
07	Working with members of this team, my unique skills and talents are valued and utilized.	I teamsamarbeidet blir mine ferdigheter og egenskaper verdsatt og brukt.

References

Abedin, B., Daneshgar, F., & D'Ambra, J. (2011). Enhancing non-task sociability of asynchronous CSCL environments. *Computers & Education, 57*(4), 2535–2547.

Baer, M., & Frese, M. (2003). Innovation is not enough: Climates for initiative and psychological safety, process innovations, and firm performance. *Journal of Organizational Behavior, 24*(1), 45–68. https://doi.org/10.1002/job.179

Balacheff, N., Ludvigsen, S., De Jong, T., Lazonder, A., Barnes, S. A., & Montandon, L. (2009). *Technology-enhanced learning.* Berlin: Springer.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bradley, B. H., Postlethwaite, B. E., Klotz, A. C., Hamdani, M. R., & Brown, K. G. (2012). Reaping the benefits of task conflict in teams: The critical role of team psychological safety climate. *Journal of Applied Psychology, 97*(1), 151–158. https://doi.org/10.1037/a0024200

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361–379). Guilford Publications.

Byrne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* Routledge.

Byrne, B. M. (2013). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* psychology press.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*(2), 234–246. https://doi.org/10.1037/0021-9010.83.2.234

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods, 8*(4), 375–409. https://doi.org/10.1177/1094428105280056

Cole, C., Marhefka, J., Jablokow, K., Mohammed, S., Ritter, S., & Miller, S. (2021). What is the relationship between psychological safety and team productivity and effectiveness during concept development? An exploration in engineering design education. *Journal of Mechanical Design, 144*(11), Article 112301. https://doi.org/10.1115/1.4054874

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Psychology, 40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly, 16*(1), 149–167. https://doi.org/10.1016/j.leaqua.2004.09.009

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

Edmondson, A. (2004). Psychological safety, trust, and learning in organizations: A group-level lens. In R. M. Kramer, & K. S. Cook (Eds.), *Trust and distrust in organizations: Dilemmas and approaches* (pp. 239–272). Russell Sage Foundation.

Edmondson, A. C., & Daley, G. (2020). How to foster psychological safety in virtual meetings. *Harvard Business Review, 25*.

Edmondson, A. C., & Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 23–43. https://doi.org/10.1146/annurev-orgpsych-031413-091305

Farnell, T., Skledar Matijevic, A., & Scukanec Smith, N. (2021). *The impact of COVID-19 on higher education: A review of emerging evidence. Analytical report.* European Commission. Available from: EU Bookshop.

Fischer, R. (2015). Multilevel approaches in organizational settings: Opportunities, challenges, and implications for cross-cultural research. In F. J. R. van de Vijver, D. A. Van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 175–198). Psychology Press.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fleischmann, C., Seeber, I., Cardon, P., & Aritz, J. (2023). Fostering psychological safety in global virtual teams: The role of reminder nudges and team-based interventions. In *Hawaii international conference on system sciences, 2023, Hawaii, USA* [Paper presentation].

Fleiss, J. (1986). *The design and analysis of clinical experiments.* Wiley.

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research, 103*(2), 231–265. https://doi.org/10.1007/s11205-011-9844-3

Fransen, J., Kirschner, P. A., & Erkens, G. (2011). Mediating team effectiveness in the context of collaborative learning: The importance of team and task awareness. *Computers in Human Behavior, 27*(3), 1103–1113.

Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A., & Vracheva, V. (2017). Psychological safety: A meta-analytic review and extension. *Personnel Psychology, 70*(1), 113–165.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91. https://doi.org/10.1037/a0032138

Gibson, C. B., & Gibbs, J. L. (2006). Unpacking the concept of virtuality: The effects of geographic dispersion, electronic dependence, dynamic structure, and national diversity on team innovation. *Administrative Science Quarterly, 51*(3), 451–495. https://doi.org/10.2189/asqu.51.3.451

Glikson, E., & Erez, M. (2020). The emergence of a communication climate in global virtual teams. *Journal of World Business, 55*(6), Article 101001.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265. https://doi.org/10.1037/0033-2909.103.2.265

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Statistical analysis* (pp. 281–294). Oxford University Press.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166–184.

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research & Development, 68*(2), 783–805. https://doi.org/10.1007/s11423-019-09729-5

Jebb, A. T., Tay, L., Ng, V., & Woo, S. (2017). Construct validation in multilevel studies. In S. E. Humphrey, & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 253–278). American Psychological Association. https://doi.org/10.1037/0000115-012.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*(5), 365–379.

Johnson, S. D., Suriya, C., Yoon, S. W., Berrett, J. V., & La Fleur, J. (2002). Team development and group processes of virtual learning teams. *Computers & Education, 39*(4), 379–393. https://doi.org/10.1016/S0360-1315(02)00074-X

Kessel, M., Kratzer, J., & Schultz, C. (2012). Psychological safety, knowledge sharing, and creative performance in healthcare teams. *Creativity and Innovation Management, 21*(2), 147–157. https://doi.org/10.1111/J.1467-8691.2012.00635.X

Kirkman, B. L., Cordery, J. L., Mathieu, J., Rosen, B., & Kukenberger, M. (2013). Global organizational communities of practice: The effects of nationality diversity, psychological safety, and media richness on community performance. *Human Relations, 66*(3), 333–362. https://doi.org/10.1177/0018726712464076

Kyriazos, T. A. (2019). Applied psychometrics: The modeling possibilities of multilevel confirmatory factor analysis (MLV CFA). *Psychology, 10*(6), 777–798. https://doi.org/10.4236/psych.2019.106051

Lee, H. (2021). Changes in workplace practices during the COVID-19 pandemic: The roles of emotion, psychological safety and organisation support. *Journal of Organizational Effectiveness: People and Performance, 8*(1), 97–128.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Marra, A., Buonanno, P., Vargas, M., Iacovazzo, C., Ely, E. W., & Servillo, G. (2020). How COVID-19 pandemic changed our communication with families: Losing nonverbal cues. *Critical Care, 24*, 1–2.

McLeod, E., & Gupta, S. (2023). The role of psychological safety in enhancing medical students' engagement in online synchronous learning. *Medical science educator, 33*(2), 423–430.

Miller, S., Marhefka, J., Heininger, K., Jablokow, K., Mohammed, S., & Ritter, S. (2019). The trajectory of psychological safety in engineering teams: A longitudinal exploration in engineering design education [proceedings paper]. In *International design engineering technical conferences and computers and information in engineering conference. Volume 7:31st international conference on design theory and methodology*. California, USA: Anaheim. https://doi.org/10.1115/DETC2019-97562. V007T06A026.

Moffett, J., Little, R., Illing, J., de Carvalho Filho, M. A., & Bok, H. (2023). Establishing psychological safety in online design-thinking education: A qualitative study. *Learning Environments Research*, 1–19.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Newman, A., Donohue, R., & Eva, N. (2017). Psychological safety: A systematic review of the literature. *Human Resource Management Review, 27*(3), 521–535. https://doi.org/10.1016/j.hrmr.2017.01.001

Pérez-Mateo, M., & Guitert, M. (2012). Which social elements are visible in virtual groups? Addressing the categorization of social expressions. *Computers & Education, 58*(4), 1234–1246. https://doi.org/10.1016/j.compedu.2011.12.014

Ramalho, M. C. K., & Porto, J. B. (2021). Validity evidence of the team psychological safety survey. *Psico-USF, 26*(1), 165–176. https://doi.org/10.1590/1413-82712021260114

Schein, E. H. (1993). How can organizations learn faster? The challenge of entering the green room. *Sloan Management Review, 34*(2), 85–92.

Schein, E. H., & Bennis, W. G. (1965). *Personal and organizational change through group methods: The laboratory approach.* Wiley.

Sjølie, E., Espenes, T. C., & Buø, R. (2022). Social interaction and agency in self-organizing student teams during their transition from face-to-face to online learning. *Computers & Education, 189*, Article 104580. https://doi.org/10.1016/j.compedu.2022.104580

Sjølie, E., & van Petegem, P. (2022). Measuring the sociability of virtual learning environments for interdisciplinary student teams – a validation study. *Scandinavian Journal of Educational Research, 68*(3), 461–472. https://doi.org/10.1080/00313831.2022.2148279

Tkalich, A., Šmite, D., Andersen, N. H., & Moe, N. B. (2022). What happens to psychological safety when going remote? *IEEE Software, 41*(1), 113–122. https://doi.org/10.1109/MS.2022.3225579

Usher, M., & Barak, M. (2020). Team diversity as a predictor of innovation in team projects of face-to-face and online learners. *Computers & Education, 144*, Article 103702. https://doi.org/10.1016/j.compedu.2019.103702

Whitton, S. M., & Fletcher, R. B. (2014). The group environment Questionnaire: A multilevel confirmatory factor analysis. *Small Group Research, 45*(1), 68–88.

Zhang, Y., Fang, Y., Wei, K. K., & Chen, H. (2010). Exploring the role of psychological safety in promoting the intention to continue sharing knowledge in virtual communities. *International Journal of Information Management, 30*(5), 425–436. https://doi.org/10.1016/j.ijinfomgt.2010.02.003