Research and Practice in
Technology Enhanced Learning

**RESEARCH** **Free and Open Access**

# How to measure disagreement as a premise for learning from controversy in a social media context

Nils Malzahn [1] *, Veronica Schwarze [2], Sabrina C. Eimler [2], Farbod Aprin [1], Sarah Moder [2] and H. Ulrich Hoppe [1]

*Correspondence:
nm@rias-institute.eu
[1] Rhine-Ruhr Institute for applied System innovation e.V., Bürgerstr. 15, 47057 Duisburg, Germany
Full list of author information is available at the end of the article

**Abstract**

Learning scenarios building on disagreement in a learning group or a whole classroom are well established in modern pedagogy. In the specific tradition of collaborative learning, such approaches have been traced back to theories of socio-cognitive conflict and have been associated with argumentative learning interactions. An important premise for these types of learning scenarios is the identification of disagreement. In the spirit of learning analytics, this calls for analytic tools and mechanisms to detect and measure disagreement in learning groups.

Our mathematical analysis of several methods shows that methods of different origin are largely equivalent, only differing in the normalization factors and ensuing scaling properties. We have selected a measure that scales best and applied it to a target scenario in which learners judged types and levels of "toxicity" of social media content using an interactive tagging tool. Due restrictions imposed by the pandemic, we had to replace the originally envisaged classroom scenario by online experiments. We report on two consecutive experiments involving 42 students in the first and 89 subjects in the second instance. The results corroborate the adequacy of the measure in combination with the interactive, game-based approach to collecting judgements. We also saw that a revision of categories after the first study reduced the ambiguity. In addition to applying the disagreement measure to the learner judgements, we also assessed several personality traits, such as authoritarianism and social closeness. Regarding the dependency of the learner judgements on personality traits, we could only observe a weak influence of authoritarianism.

**Keywords:** Toxic content, Disagreement measures, Socio-cognitive conflict, CSCL

## Introduction

It is widely accepted in modern pedagogy that disagreement and controversies can stimulate learning in classrooms and other social settings. Johnson and Johnson (1979) already stated that "disagreement among students' ideas, conclusions, theories, and opinions is an important source of learning in all instructional situations" (ibid.).

In this context, the same authors identified several relevant factors and premises that facilitate learning through conflict and controversy in the classroom, including individual heterogeneity or diversity, availability of information, or perspective taking skills. The most obvious and most frequently addressed cases and areas where disagreement arises have to do with opinions and (ethical and other) judgements.

Learning through conflict and controversy resonates particularly with established practices of computer-supported collaborative learning (CSCL). There is a variety of approaches that support the organization and orchestration of CSCL scenarios. Among these, CSCL scripting (cf. Fischer et al., 2007) aims at inducing an explicit process structure on the group activity, often using prompts and stimuli in the learning medium. The "jigsaw" method fosters knowledge exchange and knowledge building in a group by inducing a certain distribution of knowledge. Several approaches to learning driven by argumentation (Andriessen et al., 2003; Jonassen & Kim, 2010) are based on role models that guide the interaction and relationships between multiple parties that take part in the argumentation.

In this group, "ArgueGraph" (Jermann & Dillenbourg, 1999) is an example that combines a process scripting approach with role-assignments in a dyadic argumentative context. The interacting dyads are selected based on contrasting prior opinions on the subject to be discussed. The design principle behind this scenario is maximizing cognitive differences with the aim of confronting and reconciling these in a collaborative social grouping.

ArgueGraph is an example of a CSCL scenario that builds on "socio-cognitive conflict" (Mugny & Doise, 1978) as a driver of shared knowledge exchange and knowledge revision. Meanwhile we have seen more examples of learning scenarios based on socio-cognitive conflict, leading also to a widening of scope in the related research questions: This includes the influence of motivational and affective factors in such scenarios (Asterhan et al., 2010). In this perspective, the conflict should be a trigger for cognitive activity and engagement, yet not emotionally destructive for the social interaction (Näykki et al., 2014). There is evidence for the claim that confrontation should be induced and regulated based on epistemic factors, such as the distribution of knowledge, whereas socially competitive constellations should be avoided for the benefit of learning (Buchs & Butera, 2004; Butera et al., 2019). Obviously, the affective dimension is particularly relevant for our target scenario.

It is a classic finding of social psychology that sound and valid minority judgements may be dominated and overruled by majorities with less grounded judgements (Asch, 1956). There is evidence for the claim that this effect can be countered by the provision of group awareness tools (Buder & Bodemer, 2008). In the experimental setting of this study, the different viewpoints, and their distribution in the group were induced. However, in open, uncontrolled learning situations such parameters must be detected. The measurement of disagreement is an important ingredient for this.

The work reported is being conducted in a multi-party project that aims at strengthening the awareness and resilience of young learners towards discriminatory and toxic effects arising from social media usage. These phenomena are of acknowledged importance and relevance in the targeted age group of junior high-school students (Schultze-Krumbholz et al., 2012). Any supportive action to improve awareness and resilience facing such threats must consider the individually different perceptions and judgements related to such phenomena in the given learner group. Once such differences have been identified they can be dealt with as a case for learning by conflict and controversy. Accordingly, the basic "pedagogical workflow" in the envisaged classroom scenario starts with an individual activity in which the learners classify given items or instances of possibly problematic social media content using different predefined categories ("cybermobbing", "hate speech", "sexism", etc.). The individual judgements are collected using a game-like application with prepared examples. The results are stored in a database that feeds into a teachers' dashboard in which the items appear ordered and grouped according to their degree of controversy. The teacher may select examples from the spectrum of items for plenary or small-group discussions, knowing about the associated level of controversy or disagreement. The scenario and its technical orchestration allow for maintaining the anonymity of the individual learners and their concrete judgements.

The learning environment that we ultimately envisage will make use of intelligent, AI-enabled "sensoring" techniques to identify the toxicity of certain social media content. In this context, intelligent, machine learning-based algorithms appear in different roles: On the one hand, they are part of the problem, e.g., when personalized targeting of information supports the creation of filter bubbles. On the other hand, they can help to detect threats and generate supportive scaffolds to counter such risks (von der Weth et al., 2020). The detection of hate speech (MacAvaney et al., 2019) is of particular interest for providers of social media platforms. However, there are indications that solutions might not be easily at hand. Especially for the case of hate speech, there is evidence that the (human) annotation and classification is unreliable to a high degree: Using a corpus of about 14k tweets related to the European refugee crisis, Ross et al. (2017) have found only low levels of agreement between human annotators. Even the provision of clear definitions did not make a significant difference. This corroborates the assumption that we need to consider and control disagreement as an inherent factor when dealing with ethics-related subjective judgements regarding the toxicity of social media content.

In the next section we investigate the potential of assessing disagreement in a given group of learners through a systematic, statistical comparison of individual judgements. We first identify several candidate measures of disagreement and analyze their mathematical characteristics. Our analysis shows that certain approaches, which appeared to be quite different due to their different origins and contexts, are practically equivalent. Due to the restrictions caused by the COVID-19 pandemic, we have not been able to implement and evaluate the intended scenario in a face-to-face classroom setting. Instead, we have collected data in two online studies with mostly entry-level higher education students using the application originally developed for a secondary school classroom scenario. This data collection corresponds to what we would have after the first phase of our classroom scenario. In absence of the ensuing group phase, we analyzed these data to check the practicality of the disagreement measurement in comparison to "external measures" (expert ratings and behavioral parameters such as individual response times). Preliminary findings based on the first study have already been reported by Malzahn et al. (2021). Here, we extend these results by a second study including insights about supposed connections to personality traits while dealing with toxic content. We also discuss the overall findings and their relevance for further applications in offline and online educational settings.

## Measures of disagreement

Regarding the measurement of disagreement, we deal with a collection of items that have been annotated individually by students from a given group. The items are Instagram-style, text-decorated images and labels selected from a given set of categories through interactive tagging. Methodologically, this means that we must compare multiple raters who rate multiple items. In this context, the actual ratings are defined on a nominal scale, i.e., without a given inherent order. This excludes the use of most "dispersion" measures from descriptive statistics and leaves only few options. Assessing disagreement is the inverse problem of determining agreement as provided by measures of inter-rater reliability. This suggests that known measures of this type could be used in the inverse way. Given that we must deal with multiple raters and a nominal scale, Fleiss' kappa (Fleiss, 1971) is a candidate in this group. We have also found one measure of disagreement that was genuinely conceived from the perspective of collaboration research (Whitworth, 2007). Finally, we have also considered Shannon's entropy measure for comparison. There is a direct correspondence of measures of agreement ($A$) with measures of disagreement ($D$). If these measures are normalized on a scale ranging from 0 to 1, this correspondence is expressed by the equation $D = 1 - A$. In this sense, dispersion and entropy are D-measures whereas inter-rater reliability is an A-measure.

### Dispersion index (*DI*)

The "dispersion index" (*DI*) is one of the few genuine statistical dispersion measures that work with nominal or categorical variables. We rely on the description and definition given by Walker (1999):

$$DI = \frac{K(N^2 - [\sum_{k=1}^{K} f_k^2])}{N^2 \cdot (K-1)}$$

N: Number of raters
K: Number of categories
$f_k$: Number of ratings (frequencies) for each category

### Fleiss' kappa (*FK*)

Fleiss (1971) kappa is a statistical measure for evaluating the reliability of agreement between a fixed number of raters when assigning certain ratings to possibly multiple items. The FK-value is normalized and measures agreement, so that *1-FK* can serve as a D-measure.

$$FK = \frac{1}{N(N-1)} \cdot \left( \left[ \sum_{k=1}^{K} f_k^2 \right] - N \right)$$

N: Number of raters
K: Number of categories
$f_k$: Number of ratings (frequencies) for each category

### Group disagreement (*GD*)

To quantify disagreement, Whitworth (2007) introduced a measure that builds up an overall disagreement value from pairwise individual disagreement values forming a "disagreement matrix" ($d_{ij}$).

The binary value $d_{ij}$ is *0* if the two raters i and j have given different ratings (or tags), otherwise it is *1* (including for the diagonal values $d_{ij}$). An individual's disagreement ($d_i$) with the rest of the group is then the sum of disagreements with each other group member, divided by the number of pairs (*n-1*):

$$d_i = \frac{1}{N(N-1)} \cdot \sum_{j=1}^{N} d_{ij}$$

N: Number of raters
$d_{ij}$: Disagreement matrix ($d_{ij}$) based on individual ratings

The overall group disagreement is then the average of the disagreement of all its members.

If all raters and ratings agree (unanimously), the value *GD* will be 0. The maximum possible value 1 of group disagreement can only be reached if there are at least as many categories as there are raters (otherwise some raters would have to coincide in their ratings). *GD* is actually a genuine measure of disagreement. To make it comparable to the other measures targeting agreement, we can move to "group agreement" *GA* defined as *1-GD*. These measures can be formulated in the same way using an "agreement matrix" ($a_{ij}$) where

$a_{ij} = 1 - d_{ij}$. Here, the $a_{ij}$ values can be grouped and summed up in terms of the frequencies per category (for reasons of space, this cannot be fully elaborated here):

$$GA = (1 - GD\ ) = \frac{1}{N(N-1)} \cdot \left( \sum_{k=1}^{K} f_k \cdot (f_k - 1) \right)$$

Given that the sum of frequencies over all categories is equal to the number of raters, i.e., $\sum_{k=1}^{K} f_k = N$, we can rewrite the above formula:

$$GA = (1 - GD\ ) = \frac{1}{N(N-1)} \cdot \left( \sum_{k=1}^{K} f_k^2 - \sum_{k=1}^{K} f_k \right) = \frac{1}{N(N-1)} \cdot \left( \sum_{k=1}^{K} f_k^2 - N \right)$$
$$= FK$$

## Entropy-based diversity index (*H*)

Diversity or disagreement in a community can also be measured by the entropy using Shannon's formula (counting only non-empty categories):

$$H = -\sum_{k=1}^{K} \frac{f_k}{N} \cdot \log \frac{f_k}{N}$$

The highest possible value for *H* is *log(N)*. Accordingly, this measure can be normalized through division by *log(N)*.

Figure 1 shows the values of disagreement resulting from the measures *DI*, *GD* (equal to *1 – FK*) and *H* (entropy) for a simple situation with six raters giving one rating each for one item. We consider *A*, *B*, *C*, *D* as possible categorical values (however not necessarily all used). The set of example ratings is *AAAAAA*, *AAAAAB*, *AAAABB*, *AAAABC*, *AAABBB*, *AABBCC*, *AAABCD*, and *AABBCD*. The values of *GD* and *DI* appear to be very similar (see Table 1). A more detailed analysis shows that they differ only in terms of the

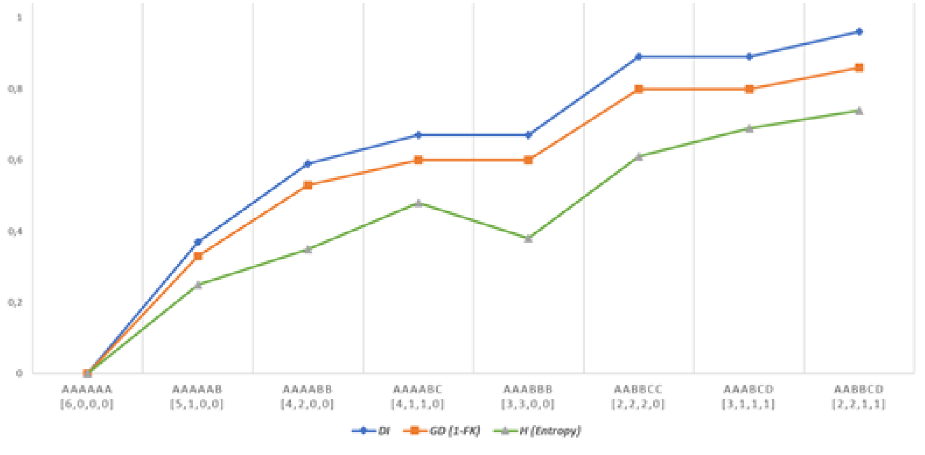


**Fig. 1** Values of disagreement for different measures (1 item, 6 raters, 4 tags)

**Table 1** Measures of disagreement / controversy with categorical judgements (nominal data)

| Group response | Pattern<br>6 out of {A, B, C, D} | Disagreement<br>1 – *pi* of Fleiss' kappa | Dispersion |
|---|---|---|---|
| Unanimous | *AAAAAA* | 0.00 | 0.00 |
| All but 1 | *AAAAAB* | 0.33 | 0.37 |
| All but 2 solid | *AAAABB* | 0.53 | 0.59 |
| All but 2 split | *AAAABC* | 0.60 | 0.67 |
| All but 3 solid | *AAABBB* | 0.60 | 0.67 |
| All but 3 split 1:2 | *AAABBC* | 0.73 | 0.82 |
| All but 3 split eq. | *AAABBC* | 0.80 | 0.89 |
| Hung group eq. | *AABBCC* | 0.80 | 0.89 |
| Max. disagreement | *AABBCD* | 0.87 | 0.96 |

normalization factor ($[N^2$ - $N]$ instead of $N^2 - \frac{1}{K}N^2$). This difference has a consequence for the possible maximum values, which is especially relevant when the number of categories is lower than the number of raters. As already noted by Whitworth (2007), the maximum value of *GD* tends to approach $(K$ - $1) / K$ for a high number of raters, which amounts to *0.5* for $K = 2$. The normalization factor of *DI* corrects for this cap in the range of values. Given that the first three measures are very similar and have been used for related purposes, we would choose one of these. We selected *DI* because of its better scaling behavior.

## Envisaged learning scenario

The original target of our design efforts was a secondary high school level classroom scenario. The targeted phenomena comprised hate speech and cyberbullying. The goal was to facilitate and support the development of strategies for counter-acting such threats on the part of the learners, while (A) creating and improving healthy social relationships among the peers, as well as (B) increasing their understanding of the social effects of toxic content and of the underlying mechanisms of propagation.

This approach is deliberately not targeted towards avoidance and external protection (e.g., by censoring, filtering, adaptation) but relies on building up understanding and reducing toxic effects by strengthening self-reflection and teaching self-protection skills and resilience. Empathy can be understood as a game changer in this context, as it is an important variable considering the individual's reactions to the observed experiences of others (Davis, 1983). As discussed by Vossen and Valkenburg (2016), e.g., this trait has a cognitive and an affective component: While the cognitive component refers to recognizing and understanding another's emotions, the affective component refers to the ability to experience another's emotions. With regard to sending toxic content on social media, this may imply that perpetrators of phenomena such as cyberbullying or online hate speech recognize the harm their behavior could cause (Hangartner et al., 2021). Empathy is also discussed as a variable related to the likelihood of showing discriminatory and hateful behavior, with victims and perpetrators of cyberbullying, for instance, showing less empathy (Schultze-Krumbholz & Scheithauer, 2009).

In addition, higher levels of state empathy (i.e., the immediate reaction of the individual to the emotional situation of others) have been shown to increase negative affect, which in turn increases willingness to respond to emotional content on social media by clicking buttons, sharing the content, or leaving a comment (Weiss & Cohen, 2019). On the other hand, in the same study, greater empathic concern (i.e., tending to experience the feelings of unfortunate people who are seen) led to less willingness to respond to social media content. The research team concludes, for instance, that compassionate adolescents resent appeals that target their negative emotions.

Closely related to empathy are socio-emotional competencies such as social awareness (Zhou & Ee, 2012). As Yang et al. (2021) discussed, social awareness is positively related to adolescents' experiences of being a victim of cyberbullying, as adolescents who appear to be more sensitive to their feelings also report being victims of cyberbullying more often. In contrast, other socio-emotional competencies such as self-management and responsible decision-making are associated with less frequent reports of cyberbullying victimization, according to Yang et al. (2021). As an explanation, they cite that young people with these qualifications are more likely to be aware of their actions online and take ethical aspects into account.

This is consistent with studies on the connection between making decisions and social closeness. As discussed by Linke (2012), social closeness refers to our social familiarity with another person. While more intimate relationships are associated with greater social closeness, low social familiarity leads to lower social closeness. Furthermore, it is explained that decisions related to oneself and members of the ingroup are evaluated differently than decisions related to outgroup members, since socially close people are of greater potential benefit.

In terms of sense of community, universal-diverse orientation is also important, defined as awareness and acceptance of similarities and differences in others (Miville et al., 2004). For instance, it can be essential for communication in a diverse social environment that people from different social and cultural backgrounds feel connected.

Regarding the requirements for interventions, authoritarianism also seems to be a relevant personality trait. Right-Wing-Authoritarianism, as a heterogenous concept, consists of three distinct components (i.e., Authoritarian Aggression, Authoritarian Submissiveness and Conventionalism), is positively associated with intolerance (Vasilopoulos & Lachat, 2018) and negatively associated with openness to experience

(Nicol & de France, 2016). If the intervention is intended to effect attitudinal change, authoritarianism could therefore be a potential challenge.

## Tool design

We have designed a serious mini game (the mobile app "SwipeIt") that allows a playful and controllable interaction with potentially toxic content items. The app was designed to be used in a classroom context to support classroom discussions around example social media items. Given that we know that the classification of social media items regarding different types of discrimination and toxic content varies between individuals, we first wanted to capture such differences and make them available as a resource for classroom discussions. In addition to triggering classroom discussions, this collection of data will also serve as a resource for further analyses using machine learning techniques.

Due to the current preferences of the target group, the app mimics the style of social media in the interaction ("swiping") and by the combination of pictorial content with embedded text or short annotations.

This also reduces language dependency, especially considering learner groups with different natural language backgrounds. In our design of the mini-game, we have been inspired by the ESP-game (Von Ahn & Dabbish, 2004).

However, since the ESP-game labels are chosen by the players (and not from a predefined set), label quality is reduced (Robertson et al., 2009).

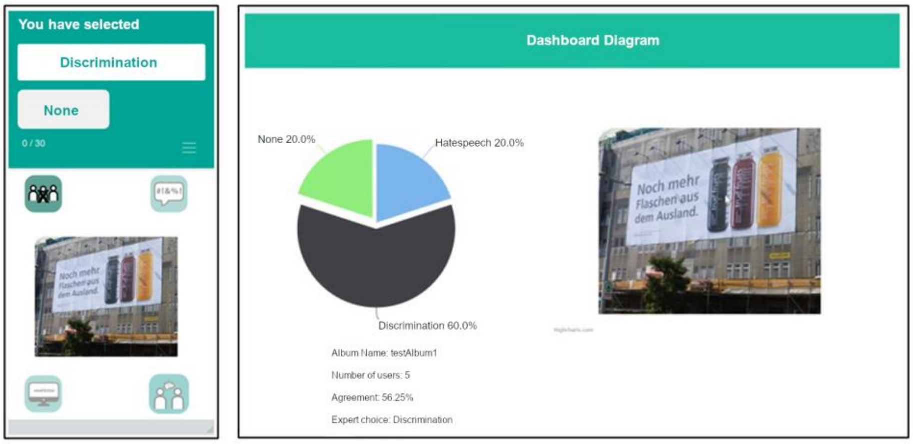
The labels to be used as tags in the mini game were selected based on semi-structured interviews with ten teachers and input from two focus groups consisting of five and three adolescents. The adolescents were between 11 and 14 years old, five female, three male. Given the age of this group, parents' consent was asked for and granted in all cases. For this group, the time spent on social media was assessed and determined to be in the range of 2 to 9 hours per week ($M$ = 4.50; $SD$ = 2.67).

The interaction with teachers and focus groups was based on existing material in the form of "reflection cards" addressing issues of racism, discrimination, and diversity (Mengis & Drücker, 2019). The members of the focus groups (adolescents) were presented with 30 pre-selected terms and were asked to choose the four most relevant ones based on their experience with social media. The selection was preceded by a short group discussion (about 5 minutes). Teachers were confronted with 39 terms written on cards as handouts. They were given the opportunity to express their thoughts on the terms and were asked to name the four most relevant terms from an imagined student's perspective. Based on the teachers' and students' choices, the four terms "verbal violence", "hate speech", "discrimination" and "cyberbullying" were selected to be used as labels for the mini-game and study.

The set of 30 images was chosen from a total of 136 images from various social media platforms such as Facebook and Instagram. All images were independently mapped to the four labels by two experts with a psychology background. The experts agreed on the same category for 67 of the 136 images. The agreement in terms of Cohen’s Kappa was $\kappa = 0.312$ ($p < .001$). The final set of 30 images contained six images from each of the previously determined four categories, plus six images classified as belonging to none of these categories. To balance this out between the 4+1 categories, four images that had not been in the selection for the experts were included. Altogether, there were five images that did not have a unanimous (or possibly no) expert rating. These images had to be excluded in the comparison of user and expert judgements but were of course included in the calculation of (dis-)agreement between users (see section Evaluation).

The SwipeIt app displays these 30 images in fixed order. Each user is asked to select the label that best describes the current image. If none of the labels is considered adequate, the user may select the option “None of the categories” shown on top of the image area and differing from the other buttons as a text on a grey area. This design should ensure that this decision is only be used as an exception, deviating from the normal operation mode.

Figure 2 shows the SwipeIt application. In the middle of the screen an image is presented to the users. In the four corners around the image, there are four buttons with icons representing the labels. In Figure 2, the user has currently selected “Discrimination” (also represented by the top left button). After deciding on a label, the user may swipe the image in any direction to see the next image as an image cannot be revisited. A label under the image indicates the progress (e.g., 1 / 30). Every interaction with any button, together with the final selection for each image, is stored in a database. Different users are distinguished by their IDs that are not connected to real names.



You have selected

Discrimination

None

0 / 30

Dashboard Diagram

Album Name: testAlbum1

Number of users: 5

Agreement: 56.25%

Expert choice: Discrimination

**Fig. 2** SwipeIt app (left) with corresponding view of the teacher dashboard (right)

## Specific research questions

Since we could only run the study without the following classroom and small group interaction, our main goal has been the validation of the agreement measure with respect to its practical usefulness to trigger controversy-based further interactions. Additionally, we were interested in a characterization of the participants in terms of emotional and social dispositions and the effectiveness of the labelling task to inspire reflection. We have analyzed and evaluated the data resulting from the usage of the SwipeIt app particularly with these questions in mind:

1. Given that the credits are not depending on answer quality, are the learner/user judgements arbitrary or do the participants actually make an effort to meaningfully and adequately characterize the examples? Indicators for this would be response times and agreement rates. It might also be that the participants tend to take the task less seriously towards the end of the completion process. This should lead to a decrease in response time and agreement rates.
2. Is answer time (time spent on one image) inter-related with “controversiality” (disagreement)?
3. For 25 out of the 30 examples, we have expert ratings that coincide in terms of a unique category assignment. How do the student classifications (agreement levels) compare to these expert ratings?
4. Given that high levels of empathy and low levels of authoritarianism predict sensitivity towards harm experienced by other people, are these personality traits reflected in the user’s tagging behavior?

## Study settings

Due to COVID-19 restrictions, the originally planned face-to-face classroom scenario with teenagers had to be adapted to an online scenario (see Figure 3) with university students (mostly entry level), to avoid confrontation of young adults with toxic content without counsel by a teacher or researcher. The original scenario was planned as follows: Teenagers are situated in a computer classroom for a session of around 90 minutes, subdivided into different phases (see Figure 3). They are introduced to the overall topic, the functionality of SwipeIt and the labels used as tags in order to ensure a shared understanding of their meaning. Subsequently, they access a questionnaire (administered using SoSci Survey) using individual codes distributed by the research team that ensure a match between the questionnaire data (across different iterations) with the game’s results. The questionnaire assesses demographic data, as well as personal and affective characteristics and contains a

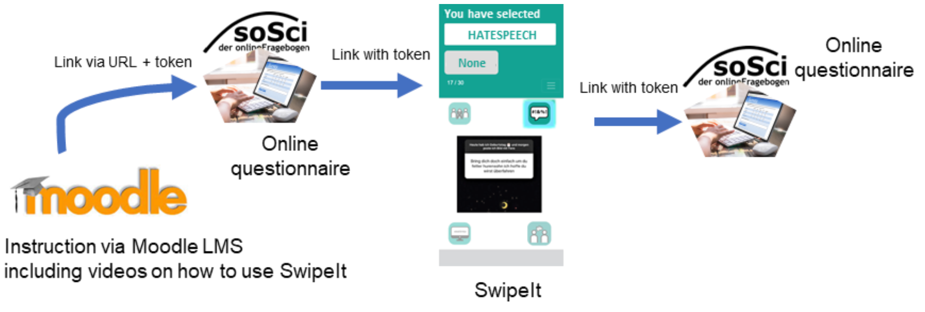


**Fig. 3** Study setting adapted to COVID-19 restrictions

link to the game that participants finish in approximately 15 minutes. Altogether each of the 30 images is tagged by every participant.

In study 1, we provided discrimination, hate speech, cyberbullying (cybermobbing) and verbal violence as labels to be used by the participants (see Figure 4 on the left). In study 2, we replaced verbal violence by sexism (see Figure 4 on the right) as the participants of study 1 reported problems with distinguishing hate speech and verbal violence. After tagging the images, a class discussion starts. Therefore, SwipeIt ranks the images according to the level of agreement for an item using the dispersion index DI. A teacher dashboard allows the moderator (e.g., teachers) to survey the spectrum of all images ranked by degree of controversy and to inspect the distribution of labels assigned by the class for each image (see Figure 2, right).

Discussing the different perceptions, e.g., using the images with highest and lowest agreement, discussing emotions, labels and elements of the content qualifying a specific label should have a sensitizing effect and inspire self-reflection.

In the adapted online-only version, the initial face-to-face teaching was replaced by introductory texts in a Moodle environment that also contained the link to the questionnaire,

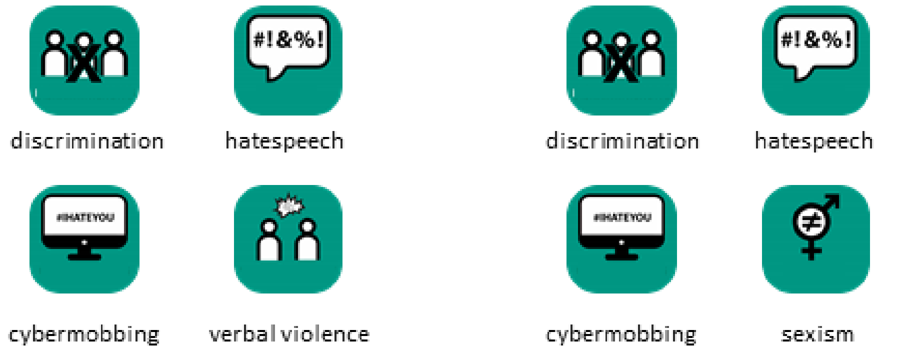


**Fig. 4** Icons used in the study 1 (left) and study 2 (right) to support the labels

using pre-defined tokens propagated between the questionnaire and SwipeIt to replace authentication procedures and allow a tracking of the answers across the different tools. With slight variations regarding the labels used in the task, the measures in the questionnaires and the procedure, we conducted two studies.

## Participants, measures, and procedure

Since we could not run the classroom scenario with secondary high school students, we recruited the participants of our online studies from beginners of a Bachelor program in Human-Computer Interaction. Although these participants were slightly older than our primary target group, we still consider them as adolescents according to Sawyer et al. (2018) who define adolescence as the period between 10 and 24 years of age. Instead of younger or early adolescents, we surveyed rather older adolescents who are becoming more independent but are still more vulnerable than adults (Sawyer et al., 2018). In addition, the students can be considered technically proficient due to their chosen course of study, so it can be assumed that participation in an online study is not a problem for them.

In both studies, participants (university students) were recruited online through a call for participation via the Moodle learning platform they were familiar with. In addition, participants in both studies received credits for participation in the studies required by their study program. To achieve the prescribed number of credits, students can choose from several studies. If they do not take part in a study or opt out, they will not suffer any disadvantage. This was pointed out in the written briefings at the beginning of the studies, which at the same time also included the respondents' declaration of consent to participate in the studies.

Before starting the interactive session that included the SwipeIt game, participants were shown a video that provided examples of the game content and functions. Here, again participants had the opportunity to decide whether to continue with the study or to opt out. Throughout the process, it was ensured that the participants could consult the research team. Email addresses for contacting, exchanging opinions or impressions were stored in the learning platform.

Study 1 data were collected from late July to late August 2020. Study 2 data were collected from mid to end of November 2020. Details about participants and questionnaires are given in the following sections.

### Study 1

The first study run (SR1) comprised 45 participants from which 42 students were included in the data analysis (three participants had missing response to questions regarding authoritarianism). The remaining participants had a mean age of 21.76 (range: 19-30; *SD* = 3.05) years; 36 participants were male and six were female. Most of them were students in their second semester (81%). German was with 69% the most common first language, followed by Turkish (19%). On a 5-point Likert scale ranging from 1 “never” to 5 “daily”, the usage of seven social media channels was assessed as familiarity with social media channels. As a result, YouTube (*M* = 4.69; *SD* = .64), Instagram (*M* = 3.81; *SD* = 1.57) and Snapchat (*M* = 2.57; *SD* = 1.68) were used most frequently, ahead of Twitter (*M* = 2.55; *SD* = 1.60), Facebook (*M* = 2.02; *SD* = 1.24), Pinterest (*M* = 1.43; *SD* = 1.01) and Tinder (*M* = 1.21; *SD* = .57).

Besides demographics and social media usage, empathy as a game changer (as outlined before) was measured by the German version of the Basic Empathy Scale (BES) by Heynen et al. (2016). The two components of empathy, affective empathy (6 items) and cognitive empathy (6 items), were both measured with a 5-point Likert scale ranging from 1 “do not agree” to 5 “fully agree”. Moreover, participants were asked to rate the social closeness to their fellow students by selecting one out of six pairs of increasingly overlapping circles. For this, we used a modified version of the Inclusion of Other in the Self (IOS) scale by Aron et al. (1992).

Further psychological traits were analyzed by adding the following two measurement instruments. The KSA-3 (Beierlein et al., 2014), a German adaptation based in part on Altemeyer’s (1981) Right-Wing-Authoritarianism scale, measured via the subscales Authoritarian Aggression, Authoritarian Submissiveness and Conventionalism. To get an overall scale score, the three subscale scores were added up and divided by three. The Miville-Guzman Universality-Diversity Scale (M-GUDS), consisting of the subscales Diversity of Contact, Relativistic Appreciation and Comfort with Differences, assessed the student’s universal-diverse orientation (Miville et al., 2004). All items (9 items on the KSA-3, 15 items on the M-GUDS) on both scales were rated in an adapted version using a 5-point Likert scale (ranging from 1 “do not agree” to 5 “fully agree”) instead of the 6-point scale used in the original version to harmonize with the other scales used in the study.

**Table 2** Concepts measured in study 1

| Measured concept | Scale | Dimensions (number of items) |
|---|---|---|
| Empathy | Basic Empathy Scale (BES) | Affective Empathy (6 items)<br>Cognitive Empathy (6 items) |
| Authoritarianism | KSA-3 | Authoritarian Aggression (3 items)<br>Authoritarian Submissiveness (3 items)<br>Conventionalism (3 items)<br>Authoritarianism total (overall scale score) |
| Universal-Diverse Orientation | The Miville-Guzman Universality-Diversity Scale (M-GUDS) | Diversity of Contact (5 items)<br>Relativistic Appreciation (5 items)<br>Comfort with Differences (5 items) |
| Social Closeness | Inclusion of Other in the Self scale (IOS) | Selecting one out of six pairs of increasingly overlapping circles |

None of the questions could be skipped when filling out the questionnaire, instead they had to be answered. Which concepts were measured with which scales in study 1 is summarized in Table 2.

After the students watched a video on how the game is proceeded, they played SwipeIt and labeled the images along the lines of discrimination, hate speech, cyberbullying (cybermobbing) and verbal violence. To mimic the classroom scenario as close as possible, a full cycle involves linking the participant back from SwipeIt to the questionnaire, to assess additional items considering emotional reactions and allowing reflection on the content (see Figure 3). However, since we focused on the agreement measures, the second part of the questionnaire was dropped from the first study.

## Study 2

The second study run (SR2) comprised 101 participants from which 89 (72 males, 16 females, 1 no indication) were included in the data analysis (e.g., two cases had missing response to questions regarding authoritarianism, three were too fast for reasonable answering). The participants had a mean age of 21.83 (range: 17-46; *SD* = 4.79) years. Most of them were first semester students (83.1%), the others were students of third (14.6%) or fifth semester (2.2%). As in study 1, the first language for most participants was German (69.7%), followed by Turkish (15.7%).

Compared to the first survey, the usage of social media channels was also assessed on a 5-point Likert scale ranging from 1 “never” to 5 “daily”. However, three additional social media channels (i.e., TikTok, Twitch, WhatsApp) were considered this time, resulting in the following preferences of the students: WhatsApp (*M* = 4.82; *SD* = .63), YouTube (*M* = 4.78; *SD* = .62) and Instagram (*M* = 3.65; *SD* = 1.65) were used most frequently, ahead of Twitch (*M* = 2.88; *SD* = 1.41), Snapchat (*M* = 2.69; *SD* = 1.80), Twitter (*M* = 2.17; *SD* = 1.45), TikTok (*M* = 1.87; *SD* = 1.41), Facebook (*M* = 1.63; *SD* = 1.19), Pinterest (*M* = 1.48; *SD* = 0.88) and Tinder (*M* = 1.28; *SD* = .88).

A second difference to study 1 was that participant’s empathy was not measured by using the Basic Empathy Scale (BES) by Heynen et al. (2016). Since skills such as Social Awareness are closely related to Empathy (Zhou & Ee, 2012) and we wanted to broaden the focus to personality traits, the Social Emotional Competencies Questionnaire (SECQ) was used instead. This questionnaire developed by Zhou and Ee (2012) is more comprehensive and comprised 25 items assigned to five subscales, representing the five dimensions of children’s and adolescent’s social emotion competence that are Self-awareness, Social Awareness, Self-management, Responsible Decision-Making and Relationship Management. Participants rated the items of those scales on a 5-point Likert scale ranging from 1 “not at all true of me” to 5 “very true of me” (6-point Likert scale in

**Table 3** Concepts measured in study 2

| Measured concept | Scale | Dimensions (number of items) |
|---|---|---|
| Social Emotional Competencies | Social Emotional Competencies Questionnaire (SECQ) | Self-awareness (5 items)<br>Social Awareness (5 items)<br>Self-management (5 items)<br>Responsible Decision-Making (5 items)<br>Relationship Management (5 items) |
| Authoritarianism | KSA-3 | Authoritarian Aggression (3 items)<br>Authoritarian Submissiveness (3 items)<br>Conventionalism (3 items)<br>Authoritarianism total (overall scale score) |
| Universal-Diverse Orientation | The Miville-Guzman Universality-Diversity Scale (M-GUDS) | Diversity of Contact (5 items)<br>Relativistic Appreciation (5 items)<br>Comfort with Differences (5 items) |
| Social Closeness | Inclusion of Other in the Self scale (IOS) | Selecting one out of six pairs of increasingly overlapping circles |
| Emotional State | Kunin Scale | One scale with sad to smiling smiley faces |

the original scale). Social Closeness, Authoritarianism (KSA-3) and the universal-diverse orientation (M-GUDS) were then measured unchanged.

Before the video with instructions on the app SwipeIt was presented and the tagging process started, the students were asked about their emotional state using 5-point Kunin scale with sad to smiling smiley faces. The tagging process itself remained unchanged, except that one label was changed. Since it appeared that “verbal violence” and “hate speech” were too similar, the label “verbal violence” was changed to “sexism” and corresponding images in two online studies were derived from literature and rated by two experts.

After the participants played SwipeIt and labeled the images along the lines of discrimination, hate speech, cyberbullying and sexism, they were redirected back to the questionnaire. The emotional state of the participants was again measured with a 5-point Kunin scale, and it was assessed which device (smartphone or mobile phone vs. laptop vs. desktop PC) participants used to complete the online survey. At the end, an open commentary field asked the students to take some minutes to reflect on their impressions about the game and the questionnaire. Additionally, they rated six items on a 5-point Likert scale (ranging from 1 “do not agree at all” to 5 “fully agree”), among others, how easy it was to use SwipeIt. The feedback collected in the open commentary field was categorized and coded as (1) “factual and focused on the conditions of the study” or (2) “self-reflective and emotional”. Notes on the procedure (e.g., “The game worked smoothly.”), on the category assignment (e.g., “Some images are difficult to assign to only one category.”) or on the questions asked (e.g., “Many questions were very similar.”) were included in the first category.

Feedback indicating a triggered reaction was summarized in the second category. This pointed to self-reflection, defined as “the inspection and evaluation of one’s thoughts, feelings and behavior" (Grant et al., 2002, p. 821), or to basic emotions like sadness, disgust and surprise (cf. Ekman, 1992).

Answers given here were, for example, "[…] thought about many things that you normally do not even notice or realize" or "Honestly, some of the statements made me tear up." The resulting codes were discussed within the research team (in part unfamiliar with the study) and adapted until consensus was reached.

If feedback referred, for example, to solving the task and at the same time contained a reference to a triggered emotional response, the entire response was categorized as "self-reflective, emotional" (e.g., "The survey and the game were quite good. It made me a little more aware of how much discrimination and cyberbullying still exists. There was nothing that was particularly difficult.").

Like the first study, it was not possible to skip any of the questions asked when completing the questionnaire, except for the question about feedback and impressions. Which concepts were measured with which scales in study 2 is summarized in Table 3.

## Evaluation results

To provide a better overview of the results, participant's personality traits of study 1 and 2 and disagreement measures are presented separately.

### Study 1

In the first study we analyzed the impact on the emotional and reflective state of the participants. It is to mention that the participants had a high level of Empathy (BES), high Universal Orientation (M-GUDS) and a tendency for low levels of authoritarianism (KSA-3). This was tested using a t-test against a test value of 3 (= middle of the scale) which showed with one exception significant deviations, i.e., tendencies to the respective scale ends. Results are shown in Table 4. When asked how they would describe their relationship with their fellow students, participants also showed a tendency for a low level. Social Closeness (*M* = 2.71, *SD* = .84) was rather low, as the theoretical mean of the scale is 3.5.

As described, five out of thirty images did not belong to any of the four categories (i.e., discrimination, hate speech, cyberbullying (cybermobbing) and verbal violence) according to the experts. Compared to the expert ratings, some participants selected the option "None of the categories" less frequently, while others selected it more frequently than six times (range: 2-25; *M* = 7.90, *SD* = 3.68). Therefore, the dataset was split at the median (8 times) into two subsets: 20 participants who used the "none" button up to seven times (range: 2-7; *M* = 5.45, *SD* = 1.67), 22 participants who used the "none" button at least eight times (range: 8-25; *M* = 10.14, *SD* = 3.60).

**Table 4** Results of t-tests regarding personality traits (study 1)

| Dependent Variable | *M* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Cognitive Empathy (BES) | 4.10 | .48 | 41 | 14.92 | <.001 |
| Affective Empathy (BES) | 3.61 | .66 | 41 | 5.97 | <.001 |
| Diversity of Contact (M-GUDS) | 3.40 | .71 | 41 | 3.69 | .001 |
| Relativistic Appreciation (M-GUDS) | 3.55 | .62 | 41 | 3.69 | <.001 |
| Comfort with Differences (M-GUDS) | 4.54 | .39 | 41 | 25.42 | <.001 |
| Authoritarian Aggression (KSA-3) | 2.56 | .95 | 41 | - 3.04 | .004 |
| Authoritarian Submissiveness (KSA-3) | 2.74 | .99 | 41 | - 1.72 | .093 |
| Conventionalism (KSA-3) | 2.36 | 1.04 | 41 | - 4.01 | <.001 |
| Authoritarianism total (KSA-3) | 2.55 | .82 | 41 | - 3.53 | .001 |

As the Levene's test showed that the variances for Authoritarian Submissiveness were not homogeneous ($F(1, 40) = 5.00$, $p = .03$), an unpaired t-test with Welch's correction was used to compare the two groups. Results indicate that both groups differed significantly in their Authoritarian Submissiveness ($t(33.12) = -2.423$, $p = .021$, $d = .75$) with participants using the "none" button less often ($M = 2.37$, $SD = 1.10$) showing a lower Authoritarian Submissiveness than participants using the "none" button more often ($M = 3.08$, $SD = .75$). In line with this, an unpaired t-test showed a significant difference in the total score for authoritarianism ($t(40) = -2.525$, $p = .016$, $d = .78$) between the two groups. Participants using the "none" button less often ($M = 2.23$, $SD = .88$) had a lower mean on the overall scale of authoritarianism than participants using the "none" button more often ($M = 2.84$, $SD = .67$). For Conventionalism ($t(40) = -1.771$, $p = .084$), Authoritarian Aggression ($t(40) = -1.947$, $p = .059$), cognitive empathy ($t(40) = .061$, $p = .95$) and affective empathy ($t(40) = .513$, $p = .61$) no significant differences were found between the two groups.

## Study 2

The participants had a high level of socio-emotional competencies (SECQ), high Universal Orientation (M-GUDS), a tendency for low levels of authoritarianism (KSA-3) and showed a tendency for a low level of social closeness towards their fellow students. This was also tested using a t-test against a test value of 3 (= middle of the scale) which showed significant deviations. Results are shown in Table 5. Social Closeness ($M = 2.74$, $SD = 1.02$) was rather low, as the theoretical mean of the scale is 3.5.

The comparison of the emotional state before and after SwipeIt using a paired-samples t-test indicated a negative mood change ($M_{before} = 3.45$, $SD_{before} = 1.01$; $M_{after} = 3.11$, $SD_{after} = .94$; $t(88) = 4.14$, $p < .001$).

**Table 5** Results of t-tests regarding personality traits (study 2)

| Dependent Variable | *M* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Self-awareness (SECQ) | 4.08 | .55 | 88 | 18.75 | <.001 |
| Social Awareness (SECQ) | 3.44 | .74 | 88 | 5.58 | <.001 |
| Self-management (SECQ) | 3.45 | .81 | 88 | 5.26 | <.001 |
| Responsible Decision-Making (SECQ) | 4.07 | .65 | 88 | 15.47 | <.001 |
| Relationship Management (SECQ) | 4.00 | .50 | 88 | 18.89 | <.001 |
| Diversity of Contact (M-GUDS) | 3.82 | .73 | 88 | 10.65 | <.001 |
| Relativistic Appreciation (M-GUDS) | 3.77 | .67 | 88 | 10.86 | <.001 |
| Comfort with Differences (M-GUDS) | 4.56 | .38 | 88 | 38.62 | <.001 |
| Authoritarian Aggression (KSA-3) | 2.55 | .86 | 88 | - 4.94 | <.001 |
| Authoritarian Submissiveness (KSA-3) | 2.66 | .99 | 88 | - 3.22 | <.002 |
| Conventionalism (KSA-3) | 2.33 | .94 | 88 | - 6.66 | <.001 |
| Authoritarianism total (KSA-3) | 2.52 | .68 | 88 | - 6.75 | <.001 |

Feedback from users collected using an open comment field was more often factual and related to the conditions of the study (*n* = 52) than self-reflective, emotional (*n* = 32). The open-ended question was not answered by five participants. Feedback received on a 5-point Likert scale, moreover, was that the items of the questionnaire were rather easy to answer, assigning the labels to the images was rather easy, such or similar images tend not to appear frequently in the student's social media channels, SwipeIt was easy to handle, but the images were hard to read. Results are presented in Table 6. To complete the survey, 51.7% of the students used a desktop PC, 46.1% used a laptop and only 2.2% a smartphone or mobile phone.

Like in study 1, five out of thirty images did not belong to any of the four categories (i.e., discrimination, hate speech, cyberbullying/cybermobbing and sexism) according to the experts. Again, compared to the expert rating, some participants selected the option "None of the categories" less frequently, while others selected it more frequently than six times (range: 0-18; *M* = 7.60, *SD* = 3.36). Therefore, the dataset of the second study was also split at the median (eight times) into two subsets: 40 participants who used the "none"

**Table 6** Evaluation of the SwipeIt experience (on a 5-point Likert scale; ranging from 1 "do not agree at all" to 5 "fully agree")

| Questionnaire item | *M* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| The questions in the questionnaire before the SwipeIt-game were easy to answer. | 3.60 | .88 | 88 | 6.42 | <.001 |
| Assigning the labels to the images in SwipeIt was easy. | 3.37 | 1.11 | 88 | 3.15 | .002 |
| Such or similar images appear frequently in my social media channels. | 2.39 | 1.41 | 88 | - 4.06 | <.001 |
| I had a good handle on SwipeIt. | 4.48 | .85 | 88 | 16.38 | <.001 |
| The pictures were hard to read. | 3.73 | 1.19 | 88 | 5.82 | <.001 |

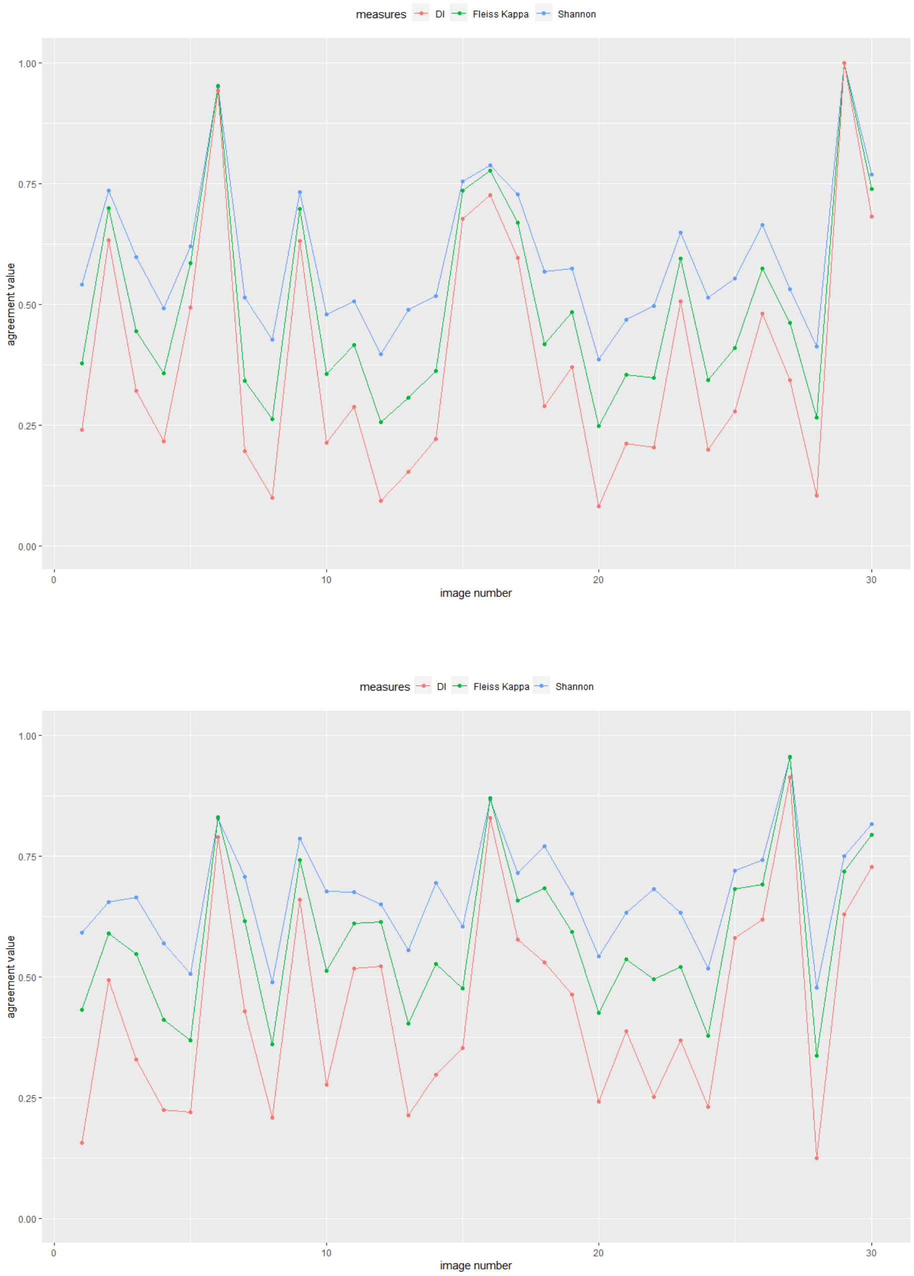


**Fig. 5** Comparison of disagreement measures for SR1 (top) and SR2 (bottom)

button up to seven times (range: 0-7; *M* = 4.77, *SD* = 2.12), 49 participants who used the "none" button at least eight times (range: 8-18; *M* = 9.90, *SD* = 2.25).

The group comparisons were performed using unpaired t-tests, yielded no significant difference for Authoritarian Submissiveness (*t*(87) = .676, *p* = .50), Authoritarian Aggression (*t*(87) = -.335, *p* = .74), Conventionalism (*t*(87) = .750, *p* = .46), Authoritarianism as an overall dimension (*t*(87) = .536, *p* = .59), Self-awareness(*t*(87) = -1.628, *p* = .11), Social Awareness (*t*(87) = -1.90, *p* = .06), Self-management (*t*(87) = -.753, *p* = .45), Responsible Decision-Making (*t*(87) = -1.577, *p* = .12) or Relationship Management (*t*(87) = -.778, *p* = .44).

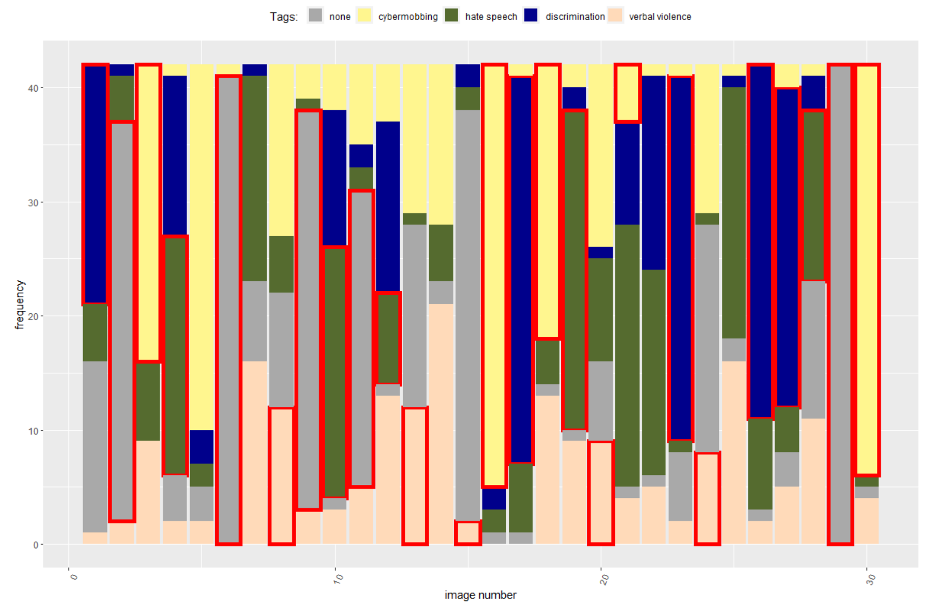


**Fig. 6** Overview of ratings for the 30 images (expert tags indicated by red frame) from SR1

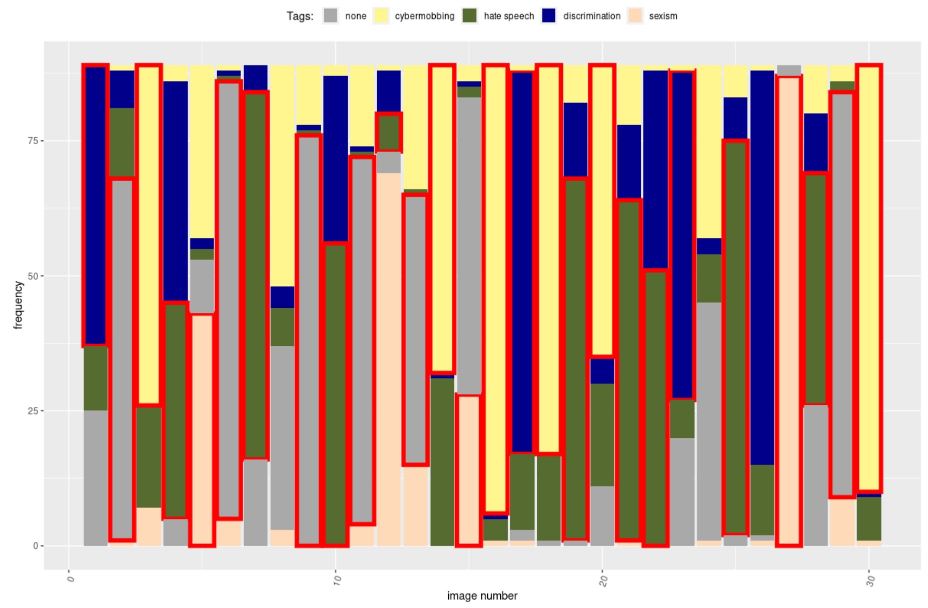


**Fig. 7** Overview of ratings for the 30 images (expert tags indicated by red frame) from SR2

## Disagreement measures

For the data analysis we have used the database with user ratings, combined expert ratings (one per image, based on consensus among two experts), and time stamps of the user actions. From the time stamps we have calculated answer times per image and user. We have also calculated disagreement (DI) per image (see Figure 5).

Figures 6 and 7 show the overall distribution of labels over the 30 items for the two studies. The items are arranged in the order of presentation with the expert ratings (tags) marked by a red frame. The grey bars stand for the neutral label "none of these". For five items in the first study there was no agreement among the experts so that no expert rating was assigned. Figure 7 shows the result for the second study where three images had no agreement in the expert ratings. In more detail, we have compared two variables: the agreement of user ratings with the expert tagging (if available), i.e., the fraction of user tags that coincide with the expert tag, and the agreement between the participant ratings measured by 1 – DI. We found a Pearson correlation of $r = .71$ ($p < .0001$) between these two parameters for the first study and $r = .71$ ($p < .0001$) for the second one. Of course, a high agreement with the expert rating would necessarily go along with a high (yet possibly smaller) agreement between user ratings, yet not necessarily vice versa. Practically, this implies that we may rely on inter-user agreement even if we do not have expert judgements as a ground truth. It also indicates that the user judgements are not just based on effortless, arbitrary guessing.

Although we have changed the label "verbal violence" (SR1) to "sexism" in SR2 Figure 8 indicates that the overall judgement of the images has not changed much from study 1 to study 2, which is also consistent with the qualitative impression gained from the (visual) comparison of Figure 5 (top) and Figure 5 (bottom). However, the average agreement among all raters and images has increased (arithmetic mean agreement in SR1: .383; arithmetic mean agreement in SR2: .480). This corroborates the assumption that the introduction of the label "sexism" reduced the ambiguity between *verbal violence* and *hate speech*. It also underlines the value of the disagreement measure for quantifying properties of the distribution of learner judgements.

Regarding answer time and disagreement (DI), there was no significant correlation ($r = .14$, $p = .5$) in the first study. This rules out the possibility of using answer time as an indicator for controversiality (here, mediated by individual insecurity). To capture sequence effects in the progression through the images we have correlated the image numbers (steps) with disagreement (study 1: $r = -.10$, $p = .62$; study 2: $r = -.29$, $p = .14$) and answer time (study 1: $r = -.14$, $p = .50$). This result indicates that there is no significant deterioration of the rating behavior when progressing through the sequence of items. This corroborates the adequacy of using the game-based scenario to elicit the judgements. There

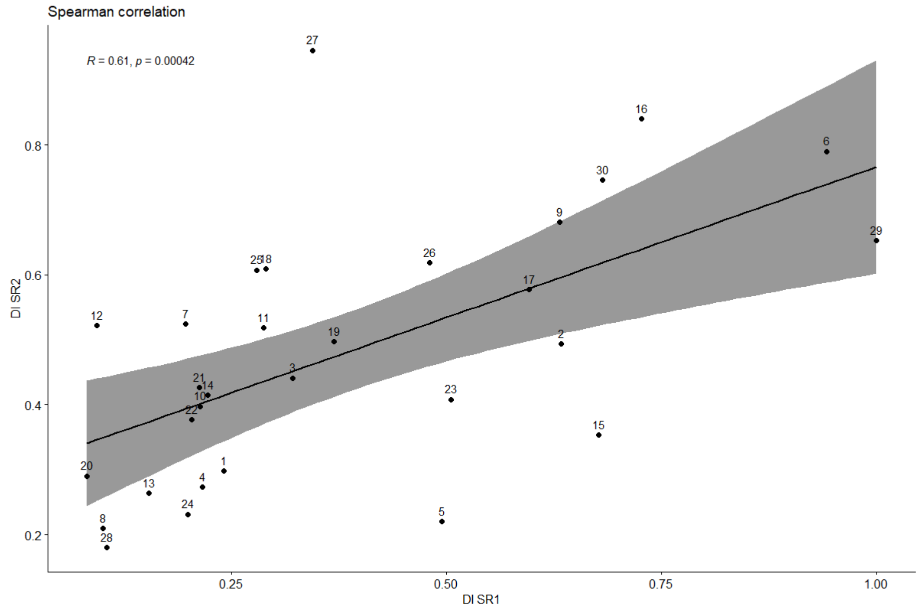


**Fig. 8** Spearman correlation of dispersion indices between SR1 and SR2

is a non-significant tendency of decreasing answer times, which can be plausibly explained by a procedural learning effect or routinization in handling the app.

## Discussion

Participants in both studies showed high levels of universal orientation, indicating that they were aware of and accepted similarities and differences in others (Miville et al., 2004). Not only did they show interest in diverse social and cultural activities, but they seem to value diversity and to be comfortable with differences in others. In addition, participants in both studies showed a tendency toward a low social closeness with their fellow students. Given that social closeness and decision-making are related (Linke, 2012), it is likely that students participated in the study without calculus. Both aspects suggest that the participants evaluated the images honestly and from their personal point of view.

Since empathy is an important variable in interpersonal perception and sensitivity to harm experienced by others (Davis, 1983; Hangartner et al., 2021), this should have been evident in the subjects’ tagging behavior. Thus, one group used the “none” button more frequently (at least eight times) compared to the other group (up to seven times). However, participants in study 1 showed high levels of cognitive and affective empathy, but the two groups did not differ significantly.

A similar result was found for socio-emotional competencies. Given that socio-emotional competencies such as self-management and responsible decision-making are linked to a higher awareness of actions online and the consideration of ethical aspects (Yang et al., 2021), this should have been reflected in a different tagging behavior of the participants as well. However, the two groups also did not differ in terms of self-awareness, social awareness, self-management, responsible decision-making, or relationship management by using the “none” button less frequently or more frequently.

At least in the first study, we found that participants who labeled fewer images than average and thus used the “none” button more often had higher levels of authoritarianism than participants who labeled more images than average and thus used the “none” button less often. However, this could only be shown for the subscale Authoritarian Submissiveness and the total score for authoritarianism. Consistent with previous literature indicating that authoritarianism is positively associated with intolerance (Vasilopoulos & Lachat, 2018) and negatively associated with openness to experience (Nicol & de France, 2016), it may be assumed that users with higher levels of authoritarianism try to avoid the labeling process in order to stay in their own mindset and to avoid thinking about phenomena such as cyberbullying, which are considered misbehavior in our society. Therefore, it seems appropriate to find further triggers to reach attitudinal changes in users with higher levels of authoritarianism. For the second study, these results could not be confirmed.

Since we changed the label “verbal violence” (SR1) to “sexism” in SR2, it is important to note that study 2 was less fraught with ambiguity. Since in both studies the “none” button was used similarly often above average, ambiguity does not appear to have been the reason for participants’ tagging behavior.

The comparison of emotional state before and after SwipeIt in SR2 showed a negative change in mood. At the same time, the “none” button was clicked an average of eight, with five images actually not assigned to any category. Following the findings of Weiss and Cohen (2019), this seems to indicate that the shown images reflected various phenomena such as cyberbullying and hate speech and elicited a negative affect, and that students were willing to respond to the images by assigning labels to them. While it is not surprising that emotional sensitivity led to reactions to sensitive content, it indicates that the labeling task stimulates emotional reactions and self-reflective behavior. This is also evident by considering the feedback in the open comments section, with 32 out of 84 responses classified as self-reflective and emotional.

In summary this confirms the assumption of research question 4 that personality traits are reflected in the user’s tagging behavior. Furthermore, the results from the comparison of the expert ratings with the students’ ratings show that most of the students agree with the experts (research question 3; cf. Figures 6 & 7). Unfortunately, the time spent on an image cannot be used as a “lightweight” indicator for controversiality, as the individual “answer time” does not predict the overall disagreement (research question 2). Although the students did not hesitate to label an image due to some kind of awareness of a conflict, they really did an effort to perform on the task meaningfully (research question 1). The analysis of study 1 clearly shows that neither the “answer time” nor the agreement depended on the position of the image in the experiment.

Some limitations must be considered when interpreting the results. First, the unbalanced gender groups need to be regarded, as in both studies fewer females than males participated. One reason for this is that computer science students were surveyed, which stereotypically tend to be male. To enhance the generalizability of the results, future studies should include more people of other genders.

The images displayed in the app represented typical social media content to model scenarios as close to reality as possible. However, on average, participants tended to agree that images of this type appeared frequently in their own social media channels. Moreover, participants tended to agree on average that the images were hard to read, although almost all the students used a desktop PC or a laptop to complete the study. Since both could be reasons for the absence of effects, images should be more closely aligned with a target group in future studies.

## Outlook

With the work reported here we intend to bring a new feature of analytics support to collaborative learning scenarios that rely on controversy and conflict. The point is to quantify the divergence (“disagreement”) between the individual learner judgements in a given group. These judgements will often be based on categorizations, i.e., they are “categorical” and cannot be compared on a numerical scale using measures such as standard deviation. Based on a mathematical analysis of several possible measures, we have first seen that they are closely related (with the exception of the entropy measure). We have selected the “dispersion index” based on its better scaling property. This measure has been built into an assessment tool that allows teachers to check and compare the controversiality of different examples in terms of the learners’ judgements. The measure can be used as an indicator and trigger of decisions (examples to select, problems to address) in the teaching situation. In this sense, our approach can serve as a tool to inform and orchestrate classroom scenarios as characterized by Johnson and Johnson (1979). The teacher interface that had been prepared for in-classroom usage provides visual support for the pedagogical decisions. Due to restrictions implied by COVID-19, we had to replace the classroom scenario by an online setting in which we could only study the individual behavior and the way this is reflected in the analytic measurement. We have clear evidence that the interplay between analytical instruments and experimental settings “works” so that we have a reasonable practical basis for further experimentation.

Although we have experienced the pandemic situation as a restricting factor, we have an interest in further supporting and extending the usage of our tools in online scenarios. In this line, we are planning to extend our online scenario with group interactions. One challenge here is the preservation of anonymity of the individual judgements. So far, our scenario does not require a combination of the internal user IDs with real identities, as long as the point is to identify the controversiality of items or artifacts. This would be different if we wanted to introduce group formation based on the characterization of users. We are currently favoring solutions that would not make use of such information, still focusing on the attribution of controversiality to the artifacts.

**Abbreviations**

BES: Basic Empathy Scale; CSCL: Computer-supported collaborative learning; DI: Dispersion index; FK: Fleiss' kappa; GD: Group disagreement; H: Entropy-based diversity index; IOS: Inclusion of Other in the Self; KSA-3: Social Closeness, Authoritarianism; M-GUDS: The Miville-Guzman Universality-Diversity Scale; SECQ: Social Emotional Competencies Questionnaire.

**Authors' contributions**

Nils Malzahn was responsible for analyzing the SwipeIt data. Farbod Aprin provided the SwipeIt application as the basic tool for the learning scenario and prepared the data needed for analysis. Ulrich Hoppe analyzed and compared the different disagreement measures based on their mathematical properties. Sarah Moder provided the questionnaire for the studies and conducted the studies with the participants. Veronica Schwarze and Sabrina Eimler analyzed the personal traits. All of them collaborated in the writing of this article.

**Authors' information**

Nils Malzahn is a senior researcher at the Rhine-Ruhr Institute for Applied System Innovation e.V. (RIAS) in Duisburg. Since he has been working with the Collide Research Group at the University Duisburg-Essen he is interested in developing intelligent and collaborative learning systems. He is also the COO of the Research Institute of Positive Computing at the Hochschule Ruhr West University of Applied Sciences in Bottrop, where he designs development tools and processes for software that aims at increasing the psychological and physical well-being of its users by being as supportive as possible for the task at hand.

Veronica Schwarze is a research assistant at the Institute of Computer Science at the University of Applied Sciences Ruhr West, working in the field of Human Factors and Gender Studies. Based on her particular interest in social diversity, she explores toxic phenomena in social media as part of the international project COURAGE.

Dr. Sabrina C. Eimler is a professor in the field of Human Factors and Gender Studies at Hochschule Ruhr West University of Applied Sciences in Bottrop, Germany. Her research covers human behavior (production and reception) in computer mediated communication among humans, especially hate practices (discrimination) in Social Media and their reduction by innovative interventions (e.g., virtual reality trainings), as well as human computer interaction, such as social robots and AI enhanced future workplaces with industrial robots.

Farbod Aprin is a research assistant that at the Rhine-Ruhr Institute for Applied System Innovation e.V. (RIAS) in Duisburg. In his PhD project, he investigates how young adults and other user groups can be supported by a virtual companion to identify and counteract toxic content in Social Media environments.

Sarah Moder is a former research assistant at the Institute of Positive Computing at the University of Applied Sciences Ruhr West. She is now working as a senior UX consultant in a German enterprise.

Dr. H. Ulrich Hoppe is an emeritus professor of "Collaborative and Learning Support Systems" of the University of Duisburg-Essen (Germany) and a fellow of the Asia-Pacific Society for Computers in Education (APSCE). His research is focused on computational techniques for analyzing and supporting collaboration, learning and knowledge building in various contexts. Currently, he is engaged as a senior researcher and head of the board of the Rhine-Ruhr Institute for Applied System Innovation e.V. (RIAS).

**Funding**

The presented research was funded by Volkswagen Foundation, Hannover (Germany).

**Availability of data and materials**

The basic data for this research (anonymized log files and questionnaire data) are accessible via Zenodo-

- Questionnaire data: https://doi.org/10.5281/zenodo.7034448
- Game / tagging data: https://doi.org/10.5281/zenodo.7034447

## Declarations

**Competing interests**

H. Ulrich Hoppe is member of the Editorial Board of RPTEL. Other than that, the authors declare that there are no competing interests.

**Author details**

[1]Rhine-Ruhr Institute for Applied System Innovation e.V., Duisburg, Germany. [2]Institute Positive Computing, Hochschule Ruhr West, University of Applied Sciences, Bottrop, Germany.

Received: 12 October 2021 Accepted: 18 August 2022
Published: 28 February 2023 (Online First: 24 October 2022)

**References**

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba Press.

Andriessen, J., Baker, M., & Suthers, D. (2003). Argumentation, computer support, and the educational context of confronting cognitions. In J. Andriessen, M. Baker & D. Suthers (Eds.), *Arguing to learn* (pp. 1–25). Springer, Dordrecht.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Asterhan, C., Schwarz, B., Butler, R., Butera, F., Darnon, C., Nokes, T., Levine, J., Belenky, D., Gadgil, S., & Sinatra, G. M. (2010). Motivation and affect in peer argumentation and socio-cognitive conflict. In *Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)*. International Society of the Learning Sciences (ISLS).

Beierlein, C., Asbrock, F., Kauff, M., & Schmidt, P. (2014). *Die Kurzskala Autoritarismus (KSA-3): Ein ökonomisches Messinstrument zur Erfassung dreier Subdimensionen autoritärer Einstellungen* (Vol. 2014/35). GESIS - Leibniz-Institut für Sozialwissenschaften.

Buchs, C., & Butera, F. (2004). Socio-cognitive conflict and the role of student interaction in learning. *New Review of Social Psychology*, *3*, 80–87.

Buder, J., & Bodemer, D. (2008). Supporting controversial CSCL discussions with augmented group awareness tools. *International Journal of Computer-Supported Collaborative Learning*, *3*(2), 123–139. https://doi.org/10.1007/s11412-008-9037-5

Butera, F., Sommet, N., & Darnon, C. (2019). Sociocognitive Conflict Regulation: How to make sense of diverging ideas. *Current Directions in Psychological Science*, *28*(2), 145–151. https://doi.org/10.1177/0963721418813986

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Ekman, P. (1992). Are there basic emotions. *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Fischer, F., Kollar, I., Mandl, H., & Haake, J. M. (2007). *Scripting computer-supported collaborative learning: Cognitive, computational and educational perspectives* (Vol. 6). Springer Science & Business Media.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Grant, A. M., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality*, *30*, 821–836. https://doi.org/10.2224/sbp.2002.30.8.821

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), e2116310118. https://doi.org/10.1073/pnas.2116310118

Heynen, E. J. E., Van der Helm, G. H. P., Stams, G. J. J., & Korebrits, A. M. (2016). Measuring empathy in a German youth prison: A validation of the German version of the Basic Empathy Scale (BES) in a sample of incarcerated juvenile offenders. *Journal of Forensic Psychology Practice*, *16*(5), 336–346. https://doi.org/10.1080/15228932.2016.1219217

Jermann, P., & Dillenbourg, P. (1999). An analysis of learner arguments in a collective learning environment. In C. M. Hoadley & J. Roschelle (Eds.), *Proceedings of the 3rd Conference on Computer-Supported Collaborative Learning (CSCL)* (pp. 265–273). International Society of the Learning Sciences.

Johnson, D. W., & Johnson, R. T. (1979). Conflict in the classroom: Controversy and learning. *Review of Educational Research*, *49*(1), 51–69. https://doi.org/10.3102/003465430490010

Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development*, *58*(4), 439–457.

Linke, L. H. (2012). Social closeness and decision making: Moral, attributive and emotional reactions to third party transgressions. *Current Psychology*, *31*, 291–312. https://doi.org/10.1007/s12144-012-9146-1

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS One*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Malzahn, N., Aprin, F., Hoppe, H. U., Eimler, S. C., & Moder, S. (2021). Measures of disagreement in learning groups as a basis for identifying and discussing controversial judgements. In C. E. Hmelo-Silver, B. De Wever & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning—CSCL 2021* (pp. 233–236). International Society of the Learning Sciences. https://doi.org/10.1145/985692.985733

Mengis, E., & Drücker, A. (2019). *Antidiskriminierung, Rassismuskritik und Diversität: 105 Reflexionskarten für die Praxis*.

Miville, M. L., Romans, J. S. C., Johnson, D., & Lone, R. (2004). Universal-diverse orientation: Linking social attitudes with wellness. *Journal of College Student Psychotherapy*, *19*(2), 61–79. https://doi.org/10.1300/j035v19n02_06

Mugny, G., & Doise, W. (1978). Socio-cognitive conflict and structure of individual and collective performances. *European Journal of Social Psychology*, *8*(2), 181–192. https://doi.org/10.1002/ejsp.2420080204

Näykki, P., Järvelä, S., Kirschner, P. A., & Järvenoja, H. (2014). Socio-emotional conflict in collaborative learning—A process-oriented case study in a higher education context. *International Journal of Educational Research*, *68*, 1–14. https://doi.org/10.1016/j.ijer.2014.07.001

Nicol, A. A., & de France, K. (2016). The Big Five's relation with the facets of Right-Wing Authoritarianism and Social Dominance Orientation. *Personality and Individual Differences*, *98*, 320–323. https://doi.org/10.1016/j.paid.2016.04.062

Robertson, S., Vojnovic, M., & Weber, I. (2009). Rethinking the ESP game. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 3937–3942). Association for Computing Machinery. https://doi.org/10.1145/1520340

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *ArXiv Preprint ArXiv:1701.08118*.

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health*, *2*(3), 223–228.

Schultze-Krumbholz, A., & Scheithauer, H. (2009). Social-behavioral correlates of cyberbullying in a German student sample. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(4), 224–226. https://doi.org/10.1027/0044-3409.217.4.224

Schultze-Krumbholz, A., Jäkel, A., Schultze, M., & Scheithauer, H. (2012). Emotional and behavioural problems in the context of cyberbullying: A longitudinal study among German adolescents. *Emotional and Behavioural Difficulties*, *17*(3–4), 329–345. https://doi.org/10.1080/13632752.2012.704317

Vasilopoulos, P., & Lachat, R. (2018). Authoritarianism and political choice in France. *Acta Politica*, *53*(4), 612–634. https://doi.org/10.1057/s41269-017-0066-9

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319–326). https://doi.org/10.1145/985692.985733

von der Weth, C., Abdul, A., Fan, S., & Kankanhalli, M. (2020). Helping users tackle algorithmic threats on social media: A multimedia research agenda. *Proceedings of the 28th ACM International Conference on Multimedia*, 4425–4434. https://doi.org/10.1145/3394171.3414692

Vossen, H. G., & Valkenburg, P. M. (2016). Do social media foster or curtail adolescents' empathy? A longitudinal study. *Computers in Human Behavior*, *63*, 118–124. https://doi.org/10.1016/j.chb.2016.05.040

Walker, J. T. (1999). *Statistics in criminal justice: Analysis and interpretation*. Jones & Bartlett Learning.

Weiss, J. K., & Cohen, E. L. (2019). Clicking for change: The role of empathy and negative affect on engagement with a charitable social media campaign. *Behaviour & Information Technology*, *38*(12), 1185–1193. https://doi.org/10.1080/0144929x.2019.1578827

Whitworth, B. (2007). Measuring disagreement. In R. A. Reynolds, R. Woods & J. D. Baker (Eds.), *Handbook of Research on Electronic Surveys and Measurements* (pp. 174–187). IGI Global.

Yang, C., Chen, C., Lin, X., & Chan, M.-K. (2021). School-wide social emotional learning and cyberbullying victimization among middle and high school students: Moderating role of school climate. *School Psychology*, *36*(2), 75–85. https://doi.org/10.1037/spq0000423

Zhou, M., & Ee, J. (2012). Development and validation of the social emotional competence questionnaire (SECQ). *The International Journal of Emotional Education*, *4*(2), 27–42.

## Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.