Perpetrators in League of Legends: Scale Development and Validation of Toxic Behavior

Bastian Kordyaka
University of Siegen
Bastian.Kordyaka@uni-siegen.de

Michael Klesel
University of Siegen
University of Twente
Michael.Klesel@uni-siegen.de

Katharina Jahn
University of Siegen
Katharina Jahn@uni-siegen.de

Abstract

Toxic behavior (TB) – a form of releasing frustration and anger in a detrimental way – is a common phenomenon in online games. Despite its importance, a validated questionnaire measuring TB is yet missing. In this paper, we apply a comprehensive procedure for scale development by using two difference sources of items. In the first one, the item pool is adapted from an existing scale. In the second one, the act frequency approach is applied to generate a pool of items. We evaluated both scales based on survey data from 380 online gamers. Both instruments are juxtaposed based on their psychometric properties. The results indicate that the adapted scale performs better in the context of our study than the scale generated from the act frequency approach and is, thus, the preferable choice. With a validated measurement scale in place, we discuss how future research can benefit from the TB scale proposed here.

1. Introduction

Within the last decade, Multiplayer Online Battle Arena (MOBA) games received an increasing attention and are still increasing in popularity. Accordingly, a considerable amount of people is playing MOBA games [13]. MOBAs can be characterized by some unique game genre elements such as a high degree of competitiveness, mastery and teamwork [13]. Due to the unique player experience, the large number of active players, and the mixture of competition and teamwork in MOBAs, related issues including aggression during games, accumulate. As a consequence, it is important to understand their roots and their consequences.

One specific aspect is Toxic behavior (TB), which already caught the interest of researchers [23]. TB is enabled through real time interaction and (mostly text based) communication between players during games and can be understood as a mental state of anger and frustration. Consequently, TB negatively affect in-game communication and contributes to a bad mood during a game. Since TB is a major driver for players' frustration it can cause several negative effects (e.g., churn of players, stress, well-being). In academia, a recent literature review identified the specific need to further explore TB [24]. From the perspective of practice, game companies (e.g., Riot Games, Blizzard, Epic) already tried to address this overall issue by teaming up in the Fair Play Alliance to fight TB and related behaviors. Their objective is to better understand underlying issues causing negative behavior, improve the player experience, and prevent the potential churn of players [31].

Despite the importance of negative behavior in online games, theory development is limited because there is no validated measurement scale for TB. This drawback hampers theory development in this domain. This drawback was also recognized by other researchers [24, 24].

With the paper at hand, we aim to close this gap and present a comprehensive development process to derive two instruments to capture TB. In specific, we use two different approaches. First, we adapt an already validated instrument from a related context ("scale adaption"). Second, we make use of a data-driven approach applying the act frequency approach ("scale building"). We select League of Legends (LoL) because it is one of the most widely played video games and is widely affected by negative behavior such as TB.

The contributions of our paper are likewise theoretical and practical. From a theoretical perspective, we contribute to existing literature by providing validated measurements. Therefore, future research has new opportunities to investigate phenomena that are related to TB. For practice, we provide further insights on the characteristics of TB, which in turn, can be used to improve the handling of TB of the gaming industry.

The paper is structured as follows. First, we introduce the related work. Next, we provide information on the methodology, present the results,

URI: https://hdl.handle.net/10125/59687
ISBN: 978-0-9981331-2-6

discussing them, and provide an outlook. We conclude the paper with a reflection of the results.

2. Related Work

2.1. Negative behavior in video games

Based on notorious theories from psychology in terms of bullying and mobbing in the real world, negative behavior and more precisely cyberbullying (CB) has become a contemporary concern in the digital world [6, 16, 17, 19]. Particularly electronic channels, without face-to-face communication, lack certain social influences, which yield in a higher perception of anonymity and deindividuation, which can lower the boundaries for TB [11, 19, 20, 29]. CB can be understood as an intentional aggressive behavior that is carried out by a group or an individual, using electronic forms of contact [27] CB can be primarily observed in social media and video games [2, 17, 19]. Regarding the latter, one specific form is toxic behavior (TB). CB is bullying online, while TB is a much more temporary behavior predominantly occurring in video games leading to frustration of players. Although both constructs overlap, they have their own merits (see Table 1).

Construct	Definition	Duration
Bullying	…an intentional behavior harassing, offending, socially excluding someone or negatively affecting someone [9]	Repeatedly
Cyber-bullying	…means an aggressive intentional behavior that is carried out by a group or an individual, using electronic forms of contact [27]	Repeatedly
Toxic behavior	…an behavior generating anger and frustration in players, harming communication, and contributing to spreading a bad mood [25]	Temporary

Table 1. Classification of negative behavior

Despite its importance, existing literature does not provide a common definition for TB. In line with previous studies, TB can be understood as a mental state of anger and frustration, which harms communication and contributes to spreading bad mood during a game [25]. Moreover, we follow the assumptions of Neto et al. [25] and understand TB as a phenomenon in the realm of video gaming, which happens when a player comes across a negative event during a match generating anger and frustration. This in turn leads to a harmed, contaminated, and disseminated toxic communication using pings and text chat. With regard to TB, examples include insulting other players, or an exaggerated usage of pings. A distinctive feature between CB and TB is temporary phenomenon in contrast to CB, which commonly emerge over a longer time period.

Several studies in Information Systems (IS) and Human Computer-Interaction (HCI) research already addressed TB, but none used quantitative self-reports from players. For instance, Blackburn et al. and Kwak et al. [2, 17] used written content and wording from players who have a tendency for a toxic behavior during a game. This is because toxic players cannot be differentiated from non-toxic players at the beginning of a game. TB rather emerges in the course of a game. Shores et al. [26] use game data to build a toxicity index contemplating a Chinese sample. They suggest that toxic players often scare away new players. They also conclude that experienced players (depending on the total amount of matched played), are more resilient towards encountered TB. Neto et al. [25] investigate communication patterns of players during a game and provide empirical evidence that they are directly linked to performance and the level of TB.

On the level of measurements, previous literature already addressed certain aspects of behaviors in video games. One noteworthy example in this regard is the Social Presence in Gaming Questionnaire (SPGQ), which describes games as social presence technologies [15]. However, a validated measurement scale for TB is missing so far. This lack is crucial since an increasing number of individuals is playing games like LoL where TB occurs regularly. Having the chance to adequately measure TB can be considered the foundation for empirical research in this domain.

2.2 About the game

Researchers already noticed the remarkably meaningfulness of League of Legends for the gaming industry and the contemporary game culture [13, 23, 24]. In terms of the content, the game is a team-based, competitive video game played in teams of five. The game is a mix of real-time strategy, tower defense, and computer roleplaying games and currently considered the most popular online game in the world of video gaming [13]. The game is characterized by its fast-paced competition and the primary goal to destruct the opposing team's nexus [7]. Within the game, the most popular game mode is ranked in which each team consists of five members, who are randomly assigned to a team with four other players on a comparable skill level. Ranked games have an average playing time of 30 to 40 minutes. Depending on the outcome of a game every player receives or loses points, which indicate his skill level. Thus, every player can move up or down in the division ladder ranging from challenger to bronze. Before each match, every player has to pick one out of more than 100 champions, which possess

different personal skills. During the course of the game, players can earn gold to buy items, which increase his champion's power. Therefore, a player can destroy enemies' towers, killing minions, or score kills and assists. The mixture of different champion skills and collaboration between players during a game are the most crucial factors deciding about winning or losing a game. To increase the chance of winning a game, players can use pings (signals, a player can send to his teammates with hotkeys if they want to point on something important on the game's map) and/or the chat function by default. In doing so, disagreements about playing styles or strategies (e.g., how to prioritize objects) occur excessively. which are increased from the pressure to win or lose points depending on the outcome of a game. As the outcome of the game determines how much points a player earns, those situations are further intensified. As a result, players get frustrated which leads to different degree of TB.

3. Research Methodology

3.1. Aim of the study

The aim of this study is twofold. First, we want to provide two comprehensively developed instruments to measure TB in LoL using self-reports of players conducting TB towards other players. Second, we want to illustrate and compare two different approaches ("scale adaption" and "scale building") and investigate their efficacy.

3.2. Research design

We applied a cross-sectional survey to develop two scales measuring *TB*. Therefore, we made use of qualitative (act frequency approach) and quantitative tools (covariance-based multivariate statistics and structural equation modeling) to develop, compare, and validate both scales.

3.3. Data collection and sample attributes

We utilized multiple channels to collect a sufficient amount of respondents for our study. First, we used official community boards referring to the survey link. Second, we asked gatekeepers personally to share our survey link within the communities they have access to. Third, we posted the link on social media platforms (i.e., Facebook and Reddit). Since the digital questionnaire is designed for self-selection, the participation was voluntary.

We collected data from 409 participants using an online questionnaire. After excluding 29 cases because of missing data and dubious answers (bogus items), our final sample included responses of 380 participants. The participants had an average age of 21 years ranging from 16 to 41 years (M = 21.03, SD = 3.92) and the vast majority consisted of male participants (334 males, 46 females). Participants stated that the highest academic degree achieved are high school diploma (184) and bachelor degree (101). Most participants indicated that they are students (230), that they play either on the servers of Europe West (252) or Europe North-East (102), and report a medium level of skill (248). Additionally, the majority of participants started to play LoL five years ago (174) and more than half of them (284) achieved the highest possible level of honor (level five).

4. Results

4.1. Preliminary work

Contextual embedding of TB

Since we wanted to validate the two instruments to develop at the end of the scale building procedure, we embedded TB in a theoretical framework to show its impact on relevant outcome parameters.

First, we wanted to include an alternative measurement for TB. We looked at the origins of TB in psychology, which offer different measures for the related constructs of bullying and mobbing [36]. Therefore, we postulated that the alternative measurement for TB shows a positive impact as an indicator for present validity. We adapted a single item (TB_SM) from existing literature ("How often do you criticize other players during a ranked game?") [36] and asked respondents on a scale from 1 ("not at all") to 7 ("definitely") about their accordance with the question (M = 3.67, SD = 2.02, skew = .11, kurtosis = -1.17).

Second, we looked for a construct with a negative effect on TB. We identified prosocialness as a relevant construct in this regard, which is defined as the set of voluntary actions one may adopt to help, take care of, assist, or comfort others [9]. Furthermore, it involves attentional and evaluative processes such as moral reasoning, social competence, and self-regulatory capacities and can promote the awareness of negative consequences of own behavior [18,26]. Therefore, we adapted an existing scale [9]. For this purpose, we presented participants thirteen statements (e.g. "I try to help my teammates") on a scale from 1 ("not at all") to 7 ("definitely") and asked for their accordance (M = 4.67, SD = 1.16, α = .90, skew = -.64, kurtosis =.58).

Third, we searched for a construct extending the impact of TB. Therefore, we utilized anger and aggression [1]. We assumed that the scale positively affects the level of TB. We adapted items to the context of our study [21] and asked participants for their accordance with five statements (e.g., "During a game I find it difficult to control my temper"). Participants answered on a scale ranging from 1 ("almost never") to 5 ("almost always", M = 2.40, SD = .85, α = .76, skew = -.47, kurtosis = -.03).

Fourth, we asked participants for their accordance with three bogus items (e.g. "I have never brushed my teeth.") to reveal participants who are not paying attention or respond dishonestly. Participants answered on a dichotomous scale ("correct/ incorrect). We eliminated every participant who answered one of the bogus wrong (9).

Sample split

After collecting the data, we split our dataset into two even parts. We called the first subsample A (SSA) and the second subsample B (SSB) to have the chance to validate explorative driven results on a set of different and independent points of data in the further course of our analysis [28]. To split the data, we used a random number.

To make sure, that the sample split did not include any unwanted confounds, we checked for effects of demographic and control variables between both subsamples. Therefore, we used the SSA and SSB as independent variables and the demographic and control variables (gender, age, education, level of play, experience, and honor level) as dependent variables. A series of t-tests suggested that the sample split did not lead to unwanted confounds regarding the two subsamples SSA and SSB ($p \geq .09$). Thus, we recorded that the split of our overall sample did not include any confounding effects.

4.2. Scale Adaption – TB Questionnaire

In a first step, we searched for an existing instrument of a construct closely related to TB. Looking at the roots of TB, while considering the competitive context of LoL, we selected an instrument measuring bullying in the workplace. The scale seemed appropriate since it describes negative behavior in small-groups and meets the need for an efficient measurement. The scale has already proven its psychometric properties and comprises aspects like criticizing, intentional interrupting, not answering, or insulting others [32]. The instrument contains five items and postulates a unidimensional solution. For the remainder of the paper, we call the first instrument to develop the toxic behavior questionnaire (TB_Q).

To adapt the instrument, we used the procedure of back-translation [2]. In a first step, we started with the original version of the questionnaire and asked a native speaker with expertise in the context of video games, to adapt the items to the new context of TB. In a second step, we gave the adapted items to another researcher who was familiar with the context of work and organizational psychology and asked him to (back)translate the items to the context of work. In a third step, we evaluated differences between the original and the back-translated versions of the questionnaire [2]. Besides some minor inconsistencies ("cry" was used instead of "whine") both versions showed similar results. We requested participants to indicate their agreement regarding the statement "When I get upset while playing League of Legends there is a considerable chance that I will…". Comparing the TB_Q to prior literature [e.g. 29], we notice that the derived item solution represents a wide scope of identified TB related topics (Table 2).

Item	Wording
v_TB_Q_1	...intentionally interrupt other players while they are writing.
v_TB_Q_2	... not answer another player who asked me something.
v_TB_Q_3	...hold others responsible making own mistakes.
v_TB_Q_4	...take away resources belonging to other players.
v_TB_Q_5	...insult other players.

Table 2. The wording of the *TB_Q* items

In a second step, we carried out an exploratory factor analysis to test the dimensionality of the scale.

(1) Test of the requirements on a level of items.

First, we looked at values on a level of items and used descriptive statistics to find out which variables met the necessary assumptions (skewness, kurtosis, a measure of sample adequacy) to be included in the further analysis. In detail, we followed the recommendations from West et al., who suggest that the skewness measures should be below the required threshold of |2| and the kurtosis measures do not exceed the value of |7| [48]. Additionally, we used recommendations of Tabachnick and Fidell requiring a threshold of > .60 of the measures of sampling adequacy [42].

All items showed acceptable results (Table 2) in case of skewness (≤ | 1.98|) and kurtosis (≤ | 4.87|). Item v_TB_Q_2 indicated a questionable measure of sampling adequacy (.58). After carefully inspecting its wording ("not answer another player who asked me something."), we decided that the item might be too inaccurate because in the context of LoL, questions for

which the answer helps the person who is asked and not the person who asked are frequent. Thus, the item is ambiguous as not answering such a question does not harm the questioner. Therefore, we excluded the item from our further analysis. All other items met the required values (≥ .72).

Item	Mean	SD	Skew	Kurtosis	MSA
v_TB_Q_1	*1.72*	*1.51*	*1.98*	*4.87*	*.79*
v_TB_Q_2	*3.84*	*2.12*	*.09*	*-1.28*	*.58*
v_TB_Q_3	*2.71*	*1.64*	*.75*	*-.12*	*.77*
v_TB_Q_4	*2.54*	*1.78*	*.89*	*-.23*	*.78*
v_TB_Q_5	*2.48*	*1.89*	*1.16*	*.10*	*.72*

Table 3. Descriptive statistics of the *TB_Q*

(2) *Test of the requirements to use an EFA.*

The Kaiser-Meyer-Olkin measure of sampling adequacy is .77 and Bartlett's test of sphericity suggests meaningful connections (p < .001) between the variables, both values indicated a justified application of an EFA [8].

(3) *Amount of extracted factors.*

We made use of the minimum average partial, parallel analysis, scree test, and Kaiser criterion. All criteria suggest a solution with only one factor, whereby the one-factor solution explains 57% of the initial variance.

(4) Selection of the factor analytical method.

Since we wanted to extensively explain the latent relationships, we carried out a maximum likelihood factor analysis. The maximum likelihood goodness of fit index indicates no significant difference between the empirical data and the postulated model (p = .82). Thus, the application of a maximum likelihood factor analysis seemed to be appropriate.

(5) Determination of the rotation method.

Since we extracted only one factor, we did not have to specify a specific rotation method.

(6) Assessment of the derived factor.

The solution with four items for the *TB_Q* indicated a one dimensional measurement of *TB*. All factor loadings were above .58 exceeding the required threshold of .40 (Figure 1).

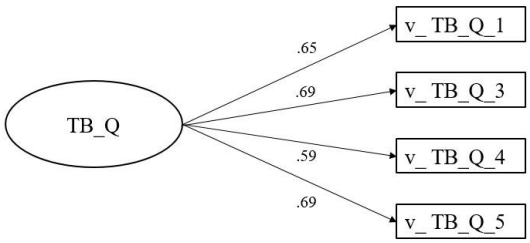

Figure 1: Exploratory analysis for the TB_Q

In a third step, we carried out a confirmatory factor analyses (CFA) to test our explorative driven results.

All items showed acceptable values regarding skew (≤|1.97|) and kurtosis (≤ |3.59|). The postulated model indicates a good overall fit in relation to the empirical data ($\chi 2$ = (2, 190) = 2.19, p = .34). Furthermore, no factor loading is below the recommended value of .40 (Figure 2). All items show highly significant values (< .001) and the share of explanation on the manifest level is at least .16 [10]. Additional fit indices confirm a good model fit (GFI = .99, RMSEA = .02). Thus, we concluded that the four-item solution of the TB_Q adequately represents a consistent construct of TB.

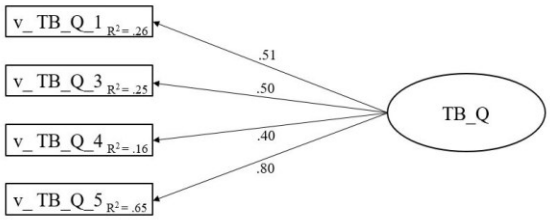

Figure 2: Confirmatory analysis for the TB_Q

Summarizing, we adapted five items of an established scale to the context of our study. Afterwards, we adjusted the scale by excluding one of the items and illustrated the unidimensional structure of the solution using an EFA. The deductive test of the scale derived in the prior step using an independent sample indicated the quantitative legitimacy of the TB_Q.

4.3. Scale Building – TB Direct Measure

In a first step, we used an empirical-driven approach to build a scale measuring TB [3]. To acquire respondents for this purpose, we collected two samples with thirty persons each. Methodologically, we followed the recommendations of the act frequency approach and proceeded in four steps. Since we used a rather direct approach, we called the second instrument to develop TB direct measure (TB_DM).

First, we asked thirty participants in an online questionnaire for their demographics and to write down their expectations and manifestations of TB regarding themselves and other players. Summarizing, the participants indicated rather homogenous answers and mentioned the aspects of cursing, insulting, whining, grieving, harassing, scamming, cheating, and using racial slurs.

Second, we took the explored aspects of the step before and tested them for their prototypicality. For this, we consulted thirty different participants in a different questionnaire. Besides demographics, we

presented them the eight aspects and asked them to evaluate their prototypicality regarding TB on a scale ranging from 1 ("not at all prototypical") to 3 ("fully prototypical"). Participants considered themselves long-time LoL players, who played around four years as an average (M = 4.24, SD = .58), most of them were male (21) and had an average age of 23 years (M = 22.76, SD = 1.78). Except twelve of them (North-America) all participants played on the EU-W Server and the average duration to answer lasted around 6 minutes (M = 5.74, SD = .62).

Third, we looked for potential confounds between the first and the second group. Thus, we executed a series of t-tests using the two groups of participants as the independent variable and the demographic variables (gender, age, origin, game experience, duration of the interview) as dependent variables. None of the five t-tests showed a significant difference between both groups ($p \geq .24$). Thus, we assumed that the grouping of participants did not include any confounding effects.

Fourth, we selected the items with a sufficient prototypicality for TB. After discussing the results, we eliminated the items v_TB_DM_7 and v_ TB_DM_8 because concerns regarding the social desirability bias in both instances TB (M = 1.40, SD = .51; M = 1.20, SD = .44). Thus, the final version of the TB_DM consisted of six items. Comparing the TB_DM items to the TB_Q items, we notice a substantial intersection on the level of content (Table 4).

Item	Wording	Mean	SD
v_ TB_ DM_1	Cursing	2.95	.22
v_TB_ DM _2	Insulting	2.90	.31
v_ TB_ DM _3	Whining	2.85	.37
v_ TB_ DM _4	Grieving	2.85	.37
v_ TB_ DM _5	Harassing	2.80	.41
v_ TB_ DM _6	Scamming	2.75	.44
v_ TB_ DM _7	Cheating	1.40	.51
v_ TB_ DM _8	Racial Slurs	1.20	.44

Table 4. Wording and prototypicality of the *TB_DM*

In a second step, we used an EFA for the *TB_DM*.

(1) Test of the requirements on a level of items.

First, we looked at descriptive values to explore whether variables met the necessary requirements to be included in the further analysis. Therefore, we used the already known thresholds one more time.

Only item *v_DM_6* ("scam someone") showed substantial violations of the necessary assumptions (skew = 6.62, kurtosis = 47.77). Due to the small discriminant power of the item, we excluded it for our subsequent analysis. All other items indicate acceptable results in case of skewness ($\leq$ |1.98|), kurtosis ($\leq$ |3.51|), and the measure of sampling adequacy ($\leq$ |.77|; Table 5).

Item	Mean	SD	Skew	Kurtosis	MSA
v_TB _DM _1	2.69	2.07	.92	-.52	.79
v_TB _DM _2	2.76	2.11	.94	-.50	.77
v_TB _DM _3	2.84	2.15	.84	-.79	.87
v_TB _DM _4	1.89	1.58	1.90	2.88	.87
v_TB _DM _5	1.83	1.71	1.98	3.51	.85
v_TB _DM _6	1.14	.72	6.62	47.77	.87

Table 5. Descriptive statistics of the *TB_DM*

(2) Test of the requirements to use an EFA

The Kaiser-Meyer-Olkin measure of sampling adequacy is .81 and Bartlett's test of sphericity indicates meaningful connections between the variables ($p < .001$). Both results suggest that the items share substantial common variance and the application of a EFA seemed suitable.

(3) Amount of extracted factors

We made use of the minimum average partial [30], parallel analysis [12], scree test [5], and Kaiser criterion [14] to identify the underlying structure of factors. All criteria suggest a solution with one factor, whereby the one-factor solution explains 65% of the initial variance.

(4) Selection of the factor analytical method

We carried out a maximum likelihood factor analysis. The maximum likelihood goodness of fit index indicates a significant difference between the empirical data and the postulated model ($p < .001$), which suggests an inaccurate fit between the empirical data and the theoretical assumptions.

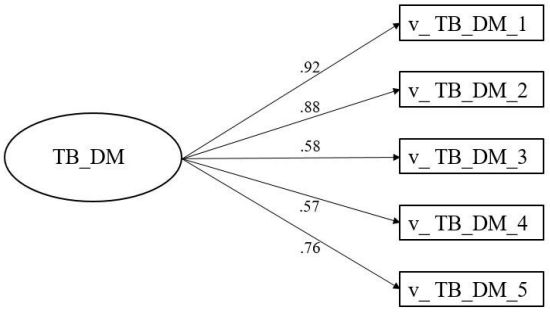

Figure 3: Exploratory analysis for the TB_DM

(5) Determination of the rotation method.

Since we extracted only one factor, we did not have to specify a specific rotation method.

(6) Assessment of the derived factor.

The derived one-factor solution for the *TB_DM* indicates a unidimensional measurement of *TB* consisting of five items (loadings < .57, Figure 3).

In a third step, we tested our explorative driven results in a confirmatory manner. Therefore, we modeled the proposed one-factor solution of TB.

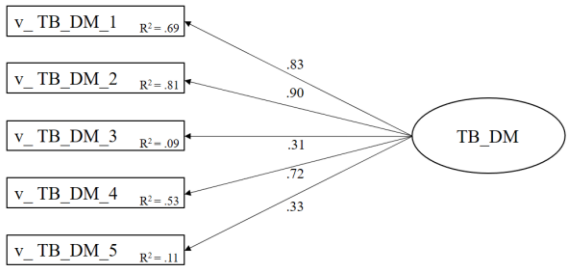

Figure 4: Confirmatory analysis for the TB_DM

Except for item v_TB_DM_6, all items showed acceptable values in terms of normality (skew ≤ |1.90|, kurtosis ≤| 6.81|). The model indicated room for improvement in relation to the empirical data (χ^2 = (5, 190) = 28.3, p < .001). All factor loadings were above ≥ .31, significant (< .01), and the share of explanation on the manifest level of items is at least ≥ .09 (Figure 4). The later indicates inaccuracies regarding the postulated model as well [10]. Additional fit indices suggested an acceptable fit (GFI = .95, RMSEA = .07). Thus, we recorded that the solution of the TB_DM provided an ambivalent picture regarding the confirmation of the explorative solution.

Summarizing, to derive items for the DM_TB we used a qualitative tool (act frequency approach) which resulted in a six item solution. Afterwards, we had to adjust the scale by excluding one of the six items. We tested the factorial structure of the solution using and CFA. The deductive test of the five factor scale derived in the prior step by using an independent sample indicated an ambivalent picture of the model fit of the DM_TB.

4.4. Scale comparison and validation

The results of both structural equation models indicated a significantly better fit between the theoretical and the empirical model for the TB_Q (χ^{2diff} = -26.16) compared to the TB_DM (χ^{2diff} = 26.16). This is consistent with the results from the fit indices of the TB_Q (GFI = .99, RMSEA = .02) and the TB_DM (GFI = .95, RMSEA = .07). Thus, we reasoned that, based on our results, the TB_Q supplies better quantitative indices compared to the TB_DM (Table 6). Although the TB_DM met the majority of required criteria's as well.

Measure	χ^2	df	p	$\chi^{2\,diff}$	MSA	RMSEA
TB_Q	2.19	2	.34	-26.16	.99	.02
TB_DM	28.35	5	.001	26.16	.95	.07

Table 6. Comparison of the *TB_Q* and the *TB_DM*

With the aim to further validate and compare our two instruments, we tested their psychometric properties (objectivity, reliability, and validity).

Objectivity.

Since the interview situation for all participants was identical approximated objectivity of our data can be assumed. To ensure objectivity, we inserted some bogus items in our questionnaire to find out whether participants answered our questions seriously and excluded those who answered in an implausible fashion.

Reliability.

First, we tested the split-half-reliability for the TB_Q and the TB_DM. Therefore, we used the two (sub)samples SSA and SSA as grouping variables. To check whether the two conditions of SSA and SSB contained significant differences regarding TB_Q values, we used an independent t-test. There was no significant difference (T (1,379) = .28, p = .60) in the scores for SSA (M = 2.36, SD = 1.29) and SSB (M = 2.29, SD = 1.24) conditions. These results indicate that the sample split did not have an effect on TB_Q, which indicates the reliability of the measurement. Following the same approach, we used the TB_DM as a dependent variable. An independent t-test shows a significant difference (T (1,379) = 13.19, p < .01) between the SSA (M = 2.40, SD = 1.54) and the SSB (M = 2.03, SD = 1.28) conditions. However, a significant Brown-Forsythe test (F = 6.55, p <.01) suggests a violation of the assumption that the group variances were statistically equal. Thus, we conducted nonparametric analysis. A significant Mann-Whitney-U test confirms the parametric results (U = 15.94, p < .05) that the SSA (Median = 201.61) condition includes marginally higher significant values than the SSB (Median = 179.39) condition. These results suggest that the sample split did have an effect on the TB_DM.

Second, we investigated the internal consistencies. Therefore, we used the two (sub)samples and the whole dataset and computed Cronbach's alpha (Table 7). The internal consistencies of both measurements regarding both subsamples indicated acceptable results (> .62). Summing up, the data regarding both TB measurements meet the necessary assumptions required for existing internal consistency.

Measure	SSA	SSB	Overall
TB_Q	.75	.62	.70
TB_DM	.86	.81	.84

Table 7. Internal consistencies of *TB* measures

Validity.

First, to ensure content validity of our measures, we asked participants at the end of our questionnaire: "Do you think the questionnaire addressed all aspects of TB? If not, what parts do you think were missing?". Although some participants provided an answer, we did not find any additional content that is not already included. Additionally, we found that the majority of aspect used in our initial definition of TB was depicted in both scales. Thus, content validity of the TB_Q and the TB_DM can be assumed.

Second, we looked at the presence of convergent and discriminant validity. Therefore, we used the computed factor scores for both instruments and the constructs of the embedding context (TB_SM, PS, AA). The convergent validity was analyzed using the average variance extracted (AVE). As the AVE is above .5 for all constructs an acceptable convergent validity is given (in the diagonal of Table 8). We assessed the discriminant validity using the Fornell-Larcker criterion. All squared correlations (values besides the diagonal in Table 8) are smaller than the corresponding construct AVE.

	TB_Q	*TB_DM*	*TB_SM*	*PS*	*AA*
TB_Q	**.52**	.71**	.63**	-.16**	.57**
TB_DM		**.57**	.61**	-.09	.61**
*TB*_SM			**1**	-.05	.57**
PS				**.51**	-.14**
AA					**.52**

Table 8. Validity indicators of TB measures

Furthermore, we looked at the correlations between the measures of TB in the embedded context. TB_Q and TB_DM show the assumed positive connections on the TB_SM ($r = .63$, $p < .001$; $r = .61$, $p < .001$) and the AA ($r = .57$, $p < .001$; $r = .61$, $p < .001$). Furthermore, both measurements correlate positively with each other ($r = .71$, $p < .001$). PS shows the postulated negative connection to the TB_Q ($r = -.16$, $p < .001$). In case of the TB_DM, PS does not reach statistical significance ($r = -.09$, $p = .08$) but shows an impact towards the assumed direction. Thus, validity regarding the direction of action is only fully detected for the TB_Q and partially for the TB_DM.

Third, the external validity of our derived instruments can be assumed since we asked real players of LoL regarding their perception in a well-known domain.

Taken together, all validity indicators (content validity, discriminant and convergent validity, external validity) showed satisfying results, which decisively strengthens the postulate of the presence of construct validity in case of both instruments.

5. Discussion

5.1. Comparison of instruments

The objective of this paper was the development of a scale to measure *TB*. Additionally, we aimed to compare two different approaches (scale adaption vs. scale building).

According to the reported model fit indices, adapting an existing instrument (*TB_Q*) showed better fit indices compared to a qualitatively built item pool (*TB_DM*). This is rather surprising since we used a standardized and widely accepted qualitative tool (act frequency approach) to develop a pool of items to measure *TB*. One possible explanation for this circumstance could be that the items of the *TB_DM* display a harsher wording on an abstract level (e.g., "take away resources" vs. "harassing") leading to a higher salience of social desirability, which might have confounded answers of some of the participants. Compared to this, the *TB_Q* rather uses specific descriptions of *TB*, which might be easier to answer and thus reducing potential confounds [25]. Another explanation for this finding might be the amount of preliminary work that guided the development of both instruments. First, we adapted an instrument with validated psychometric properties (*TB_Q*). Second, we used a qualitative tool to extract new items from scratch (*TB_DM*) [18]. Thus, the point of departure for both instruments was not similar, which could be a reason for the better fit of the scale adaption approach.

Nevertheless, during the process of developing both instruments, we had to make some decisions with inherent degrees of freedom. One instance illustrating this aspect in case of the *TB_DM* was the exclusion of items before the explorative factor analysis (precisely v_*TB_DM*_7 and v_*TB_DM*_8), where we decided to exclude items in the lower half of the middle of the distribution. Other cut-off criterions could have been used here. We wanted to take a reasonable middle road between data preservation and a strict orientation into the direction of data-driven fit. Since we used the act frequency approach in which we asked one group of participants about their expectations and manifestations

of *TB* regarding themselves and other players while using another group of participants in the subsequent step asking for the prototypically of the derived aspects. One possibility why both groups had different perceptions might be occurring differences between self-disclosure and external perception in both groups. Thus, the first group attributed the question rather to other players and the second group attributed rather internally.

Additionally, some of the internal consistencies of the *TB_Q* show rather low values (< .70). However, as *TB* comprises a variety of different strategies, we wanted to develop an instrument with high sensitivity that covers the breadth of the construct sufficiently rather than maximizing internal consistency. Item pools aimed to reflect a broad construct such as *TB* will, on average, correlate less highly with each other than will items reflecting a narrow, more tightly defined construct, because each item can only represent a smaller portion of the variance of the broad construct [32]. Our empirical finding supports this notion because both factor analyses attested the adequacy of the four-item instrument as a one-factor *TB* measure.

In the case of the *TB_DM,* the sample split had a significant effect. Although our initial analysis did not show any confounds regarding the two subsamples *SSA* and *SSB* relating to demographic variables, this result indicates an unwanted effect. We interpret this finding as an indicator that the measured content might not be as stable and influenced by other factors. This assumption is strengthened by the fluctuating distributions of the *TB_DM* (e.g. in the case of item v_*TB_DM*_6).

Looking at the validation and the embedding context of *TB*, we found several postulated connections between the two scales and relevant outcome parameters. However, in case of the *TB_DM,* the construct prosocialness only showed the expected direction but no statistical significance. We already saw that the *TB_DM* includes large(r) intragroup variances and showed a more heterogeneous picture, which might explain this finding. We understand this result as a call to expand the embedding context of *TB*. On the one hand, this could be done by adding additional constructs and on the other hand, using alternative sources of data besides self-reports inserted in an MTMM matrix [4].

5.2. Limitations and outlook

First, on a theoretical level *TB* is an emerging construct that requires further investigation. At the same time, it opens up several opportunities for future research. Thus, we suggest building a comprehensive theory which explains *TB* in future studies. This can be done by comparing different theoretical lenses capturing *TB* and by exploring new explanations merging aspects from different theories.

Second, on a methodological level, our study should be understood as an initial effort trying to develop a scale for *TB*. Thus, we documented the handling of our data extensively to provide the chance to follow our approach in detail. Further, we used self-reported values of respondents. This was intended since contemporary research lacks a questionnaire to measure *TB* using self-reports. However, future studies can try to triangulate data from different sources and explore similarities and differences between them. Furthermore, we used rather small item pools since our goal was to develop an efficient and economic measure for *TB*. Future research can try to explore additional dimensions and effects in more detail.

6. Conclusion

Since MOBAs are pervasive and increasingly played, undesired behavior such as TB affects a great number of individuals. By developing two valid scales to measure TB (TB_Q and TB_DM) for the first time, this paper contributes to an important research area and opens the door for future research related to TB. We illustrated two ways to help IS and HCI research strengthen the theoretical foundation of theorizing. Having a validated scale in place, future research is now able to quantitatively capture self-reported TB in MOBA games, which opens up a wide array of opportunities (e.g., building a theory for TB, comparing different forms of reports). Additionally, for practice (e.g., game developers, players) the usage of our scales adds value to develop better games and increase the player experience while reducing frustration during games.

8. References

[1] Berkowitz, L., "Frustration-aggression hypothesis: Examination and reformulation.", *Psychological Bulletin 106*(1), 1989, p. 59.

[2] Blackburn, J., and H. Kwak, "STFU NOOB!: Predicting crowdsourced decisions on toxic behavior in online games", ACM Press (2014), pp. 877–888.

[3] Buss, D.M., and K.H. Craik, "The act frequency approach to personality.", *Psychological Review 90*(2), 1983, pp. 105–126.

[4] Campbell, D.T., and D.W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix.", *Psychological Bulletin 56*(2), 1959, pp. 81–105.

[5] Cattell, R.B., "The scree test for the number of factors", *Multivariate Behavioral Research 1*(2), 1966, pp. 245–276.

[6] Dadvar, M., and F. De Jong, "Cyberbullying detection: a step toward a safer internet yard", *Proceedings of the 21st International Conference on World Wide Web*, ACM (2012), pp. 121–126.

[7] Donaldson, S., "Mechanics and metagame: Exploring binary expertise in league of legends", *Games and Culture 12*(5), 2017, pp. 426–444.

[8] Dziuban, C.D., and E.C. Shirkey, "When is a correlation matrix appropriate for factor analysis? Some decision rules.", *Psychological Bulletin 81*(6), 1974, pp. 358–361.

[9] Einarsen, S., *Bullying and emotional abuse in the workplace international perspectives in research and practice*, Taylor & Francis, London; New York, 2003.

[10] Hair, J.F., W.C. Black, B.J. Babin, R.E. Anderson, and R.L. Tatham, *Multivariate data analysis*, Prentice Hall Upper Saddle River, NJ, 1998.

[11] Hall, A.T., D.D. Frink, and M.R. Buckley, "An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability", *Journal of Organizational Behavior 38*(2), 2017, pp. 204–224.

[12] Horn, J.L., "A rationale and test for the number of factors in factor analysis", *Psychometrika 30*(2), 1965, pp. 179–185.

[13] Johnson, D., L.E. Nacke, and P. Wyeth, "All about that base: differing player experiences in video game genres and the unique case of moba games", *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM (2015), pp. 2265–2274.

[14] Kaiser, H.F., "The application of electronic computers to factor analysis", *Educational and Psychological Measurement 20*(1), 1960, pp. 141–151.

[15] Kort, Y.A.W.D., W.A. Ijsselsteijn, and K. Poels, "Digital games as social presence technology: Development of the social presence questionnaire", *Presence 2007 Conference*, pp. 25–27.

[16] Kowalski, R., G. W Giumetti, A. Schroeder, and M. R Lattanner, Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth, *Psychological Bulletin 140.4, 2*014, pp. 1073–1137.

[17] Kwak, H., J. Blackburn, and S. Han, "Exploring cyberbullying and other toxic behavior in team competition online games", *ACM Press (2015),* pp. 3739–3748.

[18] Leymann, H., "The content and development of mobbing at work", *European Journal of Work and Organizational Psychology 5*(2), 1996, pp. 165–184.

[19] Lowry, P.B., G.D. Moody, and S. Chatterjee, "Using IT design to prevent cyberbullying", *Journal of Management Information Systems 34*(3), 2017, pp. 863–901.

[20] Lowry, P.B., J. Zhang, C. Wang, and M. Siponen, "Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model", *Information Systems Research 27*(4), 2016, pp. 962–986.

[21] Maxwell, J.P., and A.J. Visek, "Unsanctioned aggression in rugby union: Relationships among aggressiveness, anger, athletic identity, and professionalization", *Aggressive Behavior 35*(3), 2009, pp. 237–243.

[22] Meade, A.W., and S.B. Craig, "Identifying careless responses in survey data", *Psychological Methods 17*(3), 2012, pp. 437–456.

[23] Mora-Cantallops, M., and M.-Á. Sicilia, "Exploring player experience in ranked league of legends", *Behaviour & Information Technology*, 2018, pp. 1–13.

[24] Mora-Cantallops, M., and M.-Á. Sicilia, "MOBA games: A literature review", *Entertainment Computing*, 2018.

[25] Neto, J.A.M., K.M. Yokoyama, and K. Becker, "Studying toxic behavior influence and player chat in an online video game", *ACM Press (2017)*, pp. 26–33.

[26] Shores, K.B., Y. He, K.L. Swanenburg, R. Kraut, and J. Riedl, "The identification of deviance and its impact on retention in a multiplayer game", *ACM Press*, 2014, pp. 1356–1365.

[27] Smith, P.K., J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils", *Journal of Child Psychology and Psychiatry 49*(4), 2008, pp. 376–385.

[28] Turker, D., "Measuring corporate social responsibility: A scale development study", *Journal of Business Ethics 85*(4), 2009, pp. 411–427.

[29] Vance, A., P.B. Lowry, and D. Eggett, "Using accountability to reduce access policy violations in information systems", *Journal of Management Information Systems 29*(4), 2013, pp. 263–290.

[30] Velicer, W.F., "Determining the number of components from the matrix of partial correlations", *Psychometrika 41*(3), 1976, pp. 321–327.

[31] "Fair Play Alliance", http://www.fairplayalliance.org/, accessed April 26, 2018

[32] Garthus-Niegel, S., Nübling, M., Letzel, S., Hegewald, J., Wagner, M., Wild, P. S., ... & Liebers, F, "Development of a mobbing short scale in the Gutenberg Health Study", *International archives of occupational and environmental health, 89(1)*, pp. 137-146.