Psychology of Popular Media Culture
2015, Vol. 4, No. 2, 97–111

2160-4134/15/$12.00 DOI: 10.1037/ppm0000004

Testing the Reliability and Validity of Different Measures of Violent Video Game Use in the United States, Singapore, and Germany

Robert Busching
University of Potsdam

Douglas A. Gentile
Iowa State University

Barbara Krahé and Ingrid Möller
University of Potsdam

Angeline Khoo
National Institute of Education

David A. Walsh
Mind Positive Parenting, Minneapolis, Minnesota

Craig A. Anderson
Iowa State University

To examine the potential link between violent video game play and aggressive behavior, reliable and valid measures of the level of violence in video games are required. A range of different approaches have been used in the literature, and it is not clear to what extent they converge or measure different constructs. To address this question, three large longitudinal data sets covering at least 12 months were used from the United States ($N = 1,232$), Singapore ($N = 3,024$), and Germany ($N = 1,715$). Violent content was measured through user ratings, expert ratings, and official agency ratings of individual titles as well as through expert ratings of game genres. The different measures were linked to aggressive behavior both cross-sectionally and longitudinally in all three countries and to the normative acceptance of aggression in Germany and Singapore. User ratings, expert ratings, and official agency ratings were found to be reliable. They showed substantial correlations within each culture as well as between the different cultures, indicating high convergent validity. Measures using nominations of game titles and measures using genre lists showed similar relationships with aggressive behavior and aggressive norms, both concurrently and prospectively over 12 months. Recommendations for a best practice approach to the assessment of violent content in video games are derived from the findings.

Keywords: computer games, aggression, questionnaires, measurement, media violence

In recent years, many studies have examined the effects of violent video games on aggressive thoughts, feelings, and behaviors. These studies have created a fair amount of controversy both

This article was published Online First August 26, 2013.

Robert Busching, Department of Psychology, University of Potsdam; Douglas A. Gentile, Center for the Study of Violence, Iowa State University; Barbara Krahé, and Ingrid Möller, Department of Psychology, University of Potsdam; Angeline Khoo, Psychological Studies Academic Group, National Institute of Education; David A. Walsh, Mind Positive Parenting, Minneapolis, Minnesota; Craig A. Anderson, Center for the Study of Violence, Iowa State University.

The authors gratefully acknowledge the support of Claudia Ahlert and Marianne Hannuschke. The U.S. study was supported by grants from Medica Foundation, Healthy and Active America Foundation, Cargill, and Fairview Health Services (Walsh & Gentile, PIs). The German study was supported by a grant from the German Research Foundation (Krahé, PI). The Singapore study was supported by a grant from the MOE and MDA (Khoo, PI).

Correspondence concerning this article should be addressed to Robert Busching, Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24-25, D-14476 Potsdam, Germany. E-mail: robert.busching@uni-potsdam.de

within and outside the scientific community. Some of the controversy has focused on methodological issues, such as how to measure violence in video games (Anderson et al., 2010; Ferguson, 2010). Many different approaches have been used to measure the violence levels in people's media diet, including user reports of individual games (e.g., Anderson & Dill, 2000) and ratings of game genres (e.g., Möller & Krahé, 2009). Violent content has been assessed in different ways, such as through ratings by official agencies (e.g., Kutner & Olson, 2008), experts (Wei, 2007), and users (Funk, Buchman, Jenks, & Bechtoldt, 2003). This diversity has positive and negative aspects. It is generally beneficial to the field because it allows for conceptual replication, testing the construct in a stringent way by requiring that it be robust to measurement differences (Roediger, 2012). It is potentially harmful to the field if studies using different methods find different results, which can lead to difficulties in interpreting the overall evidence. For example, it is becoming common for researchers using one method to claim that their findings fail to replicate others', and to interpret this as a lack of support for the link between violent games and aggression. However, as Roediger (2012) notes, the failure to find evidence with one paradigm does not necessarily have implications for the other paradigms. The purpose of the present article is to bring together longitudinal data sets from three continents to compare directly different methods of measuring violence in video games, testing their reliability and convergent validity as well as the predictive validity of each approach in predicting outcomes a year later.

Ratings of violent video game contents are often used to investigate the link between violent video game use and aggression-related outcome variables. Six meta-analyses, based on over 100 studies of violent video games, have been published (Anderson, 2004; Anderson & Bushman, 2001; Anderson et al., 2010; Ferguson, 2007a, 2007b; Sherry, 2001). Although they vary greatly in terms of how many studies they include, they find almost identical effect sizes for violent video games on aggressive thoughts, feelings, and behaviors (approximately $r = 0.15$ to 0.25). The empirically defined effect sizes are in the small to moderate range, and scientists often differ in their interpretations of the importance of a given effect size. This is certainly the case in this field of study, with some researchers arguing that the small-to-moderate effect sizes indicate a lack of importance and others arguing that they are highly important. The purpose of this study is not to address that debate, but to examine whether different operationalizations of violent game play converge and how they are linked to aggression-related outcome measures.

Conceptually, it is possible to separate the measurement of violent game play into three independent parts. One part is to obtain information about the media diet itself (i.e., what games users play). A second is to assess the frequency or amount of time played, and the third part is to assess the games' violent content.

The simplest approach toward measuring the games users play is to ask them directly to name their favorite or most frequently played games (e.g., *"Title of your 'most played' game,"* Anderson & Dill, 2000). This approach focuses on the game as the smallest unit, but it is also possible to shift the focus to a more general level. Rather than asking players to nominate specific games they play, they can be asked to describe the game genres they typically play. In this approach, participants are presented with a list of common game genres and asked to indicate the amount of time they usually spend playing games belonging to each genre (e.g., *"How often do you play first-person shooter computer games?,"* Richmond & Wilson, 2008).

The second part involves assessing how much time a person spends playing games, either in general, or for each specific game or genre. The third part of assessing violent game play involves a violent content judgment of either the individual game titles or the broader genres to which they may be assigned. Although there is no single "gold standard" for measuring violent content in video games, there are three common approaches: (1) to ask the participants to judge the amount of violence in a game or genre, (2) to use official game ratings, and (3) to use independent raters to assess the level of violent content in the games or genres.

A straightforward and widely used approach to measuring violent content is to ask the participants to rate the level of violence of each of their nominated games. They can either be asked to rate the global amount of violence in the game (e.g., *"how violent is the content of*

this game," Anderson & Dill, 2000), or they can be asked to rate specific behavioral aspects of violence in the game (e.g., "*how often do you kill players in this game,"* Gentile, Choo et al., 2011).

The second approach is to use the assessment of expert raters who judge the violent content of the games or genres listed by the participants. At least two types of expert ratings can be generated. First, people who know a lot about video games, such as experts from the video game industry or journalists working for gaming magazines, can be asked to rate each genre or game (e.g., Krahé & Möller, 2004). In the measures included in the present study, experts were first presented with a definition of violence and then asked to provide an independent informed opinion about the typical level of violence in a given game or game genre. A second type of expert rating can be provided by raters who are trained to identify and code specific types of violent content in clips of game sessions (e.g., Höynck, Mössle, Kleimann, Pfeiffer, & Rehbein, 2007). Although no two individual sessions of game play will be identical, it is still likely that the sessions are representative of most game play (at least for the same level of the game). Therefore, clips of game play can be coded, allowing independent raters to perform a content analysis of each clip so that interrater reliability can be established.

A third approach is to use official agency ratings. These are available in many countries, but they differ between countries. Every game released in the United States is given a rating by the Entertainment Software Ratings Board (ESRB), typically ranging from games intended for early childhood (EC), audiences of every age (E), for everyone 10 and up (E10+), for teenagers (T), for mature audiences only (M), or for adults only (AO) (Ferguson, 2011). Similarly, most games released in Germany is given a rating by the Unterhaltungssoftware Selbstkontrolle (USK), and ratings range from games intended for audiences of every age (0), for users from the age of 6 (6), uses from the age of 12 (12), users from the age of 16 (16), to mature audiences only (18). Although the two organizations base their ratings on both sexual and violent content, the USK puts more emphasis on violence than on sex (Hyman, 2005) whereas the ESRB seems to place more emphasis on sexual content (Gentile, 2008).

Each of these three different approaches has its own strengths and difficulties. Asking the user directly to assess the violent content of the game is straightforward, but could lead to problems regarding common method bias, like social desirability or implicit theories about the researchers' aims and hypotheses (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Using official agency ratings has high face validity and circumvents the problem of common method bias, but in most countries the ratings of official organizations consider aspects other than the violent content of the game (Hyman, 2005) and they provide an age-based rating rather than a content-based rating (Gentile, 2008). Furthermore, age-based ratings rest on the highly questionable assumption of a general consensus that different levels of violence are acceptable for children of different ages. A national survey of parents in the U.S. found considerable disagreement about the age at which different types of violent content are acceptable for children (Gentile, Maier, Hasson, & de Bonetti, 2011). Using expert coders has the advantage of having a clear definition of violence, allows for testing interrater reliability, as well as yielding independent assessments, avoiding the problem of common method bias. The problem with expert ratings is that their accuracy depends on the experts' knowledge and, in the case of ratings based on video clips of game play, on the representativeness of the chosen clip. Because there is no perfect measure, it is reasonable to test whether the three commonly used approaches lead to the same conclusions about the level of violent content of a given game or genre.

With the title-based and the genre-based approach, research participants' use of video games can be differentiated into violent and nonviolent game playing. In this way, it is possible to investigate the differential effect of violent and nonviolent games on aggression-related outcome variables. The General Aggression Model (GAM, Anderson & Bushman, 2002) predicts that especially video games containing at least some violence should lead to an increase in aggression-related outcomes; therefore, the use of violent games should be a better predictor of aggression-related outcomes than nonviolent game play or undifferentiated total amount of game play.

The present research was designed to contribute to the development of a generalizable best practice approach toward measuring violent video game use by examining the psychometric properties of several operationalizations of violent video game use in data sets from three countries. The aim was to provide a better understanding of the extent to which previous findings relating violent video game play to a range of outcome measures might have been affected by differences in the operationalization of violent video game use and to offer an informed recommendation about which measure(s) of violent video game play to use in future studies. Drawing on longitudinal data sets from three countries, four quality criteria were examined in this study:

(1) **Reliability**. Ratings of violent content in video games can be said to be reliable to the extent that different raters arrive at similar assessments of the violence level of a game. Intercoder agreement provides an index of reliability, enabling us to analyze the convergence of violence ratings in the form of genre ratings by experts, expert ratings of individual game titles, and user ratings of individual game titles. Establishing intercoder agreement is critical because standard indicators of reliability, such as internal consistency and test–retest reliability, are not appropriate in the case of video game play. Children may play a mix of violent and nonviolent games at any point in time and their favorite games can change over time (Anderson, Gentile, & Buckley, 2007, p .99; Gentile, Lynch, Linder, & Walsh, 2004). Therefore, to assess reliability in measures of violent video game play, it is critical to demonstrate that different sources of information yield the same conclusions regarding the level of violence in games or game genres.

(2) **Convergent validity**. Intercorrelations between the different operationalizations of violent video game use can be used to determine the degree to which different measures of violent video game content show convergence within and across the data sets. In addition, at the level of individual game titles, it is important to demonstrate that different measures of establishing violent content converge within and across the data sets from the three countries.

(3) **Construct validity or predictive validity**. A further aim was to examine how similar the different operationalization of violent video game use would be in terms of their associations with aggression-related outcome measures. To address this aim, we examined both the cross-sectional correlations between violent video game use and outcomes as well as the predictive validity of the video game measures for aggression over time. The more similar the measures are in their links with aggression, the greater confidence we can have in their validity in representing the underlying construct of exposure to violence in video games.

(4) **Discriminant validity**. Finally, it is important to demonstrate that measures of violent video game play differ from measures of nonviolent video game play as well as measures of overall gaming in their links with aggression-related outcome variables. If, as claimed for instance by the GAM, the link between video game use and aggression is driven primarily by the violent content of the games, the association between measures of violent video games and aggression should be higher than the corresponding associations of nonviolent game use or global video game use.

Thus, the present study offers a multimethod approach to the assessment of violent content in video games and the links of violent video game play with aggressive behavior and aggression-related normative beliefs, both cross-sectionally and over time. It addresses the question of how well different measures used in the empirical literature and exemplified by the three data sets converge in capturing the underlying construct of violent game content. The aim is to derive data-based recommendations for a best practice approach that may inform future studies on the effects of violent video game play on aggression. The three data sets that were selected are uniquely suitable for the purposes of this study. On the one hand, they show sufficient overlap to facilitate parallel analyses in each country for examining the convergence of measurement approaches across the different data sources, including a sample of specific video games for which violent content ratings were available from all three countries. On the other hand, they are sufficiently different to be able to complement one another and allow for a greater variety of measures of violent video game content to be used as predictors of out-

come measures related to aggression, thereby increasing the generalizability of the results.[1] To cope with the differences in the measures used in the three data sets, correlations and partial correlations were used for analyzing the data in a common metric.

Method

United States

Sample. American children were recruited between September and December 2005 from public elementary schools in two states: four schools in Lakeville, MN (population ~50,000) and six in Cedar Rapids, IA (population ~125,000). Both communities were involved in a community-, school-, and family based intervention for the prevention of childhood obesity (Gentile et al., 2009). Before participation, parents provided active written consent, and children provided assent.

A sample of 1,323 children in 3rd ($N = 430$), 4th ($N = 446$), and 5th ($N = 423$) grades returned consent forms, yielding a 65% participation rate. Forty-seven percent of participants were male (618 male, 704 female, 1 unknown) and most of the children (90%) were White, which is representative of the communities from which they were sampled. The mean age of the students at the first wave of data collection was 9.6 years ($SD = 0.9$; range 6–12 years of age). Data were collected 13 months apart. Out of 1,323 consented families, 1,288 children (97%) provided data. Of those, 1,196 children (93%) provided data at Time 1 and 1,110 children (86%) at Time 2. For 1,029 participants (80%), data were available at both Time 1 and Time 2. Data were also gathered from the children's teachers. The study was approved by the University of Minnesota Institutional Review Board in accordance with the Declaration of Helsinki and the American Psychological Association's 'Ethical Principles of Psychologists and Code of Conduct.'

Measures.

Exposure to violence in games. Three measures of violent video game content were collected. First, participants listed their three favorite games and rated how violent each of the games was. Violence ratings were collected on a 4-point scale ranging from (1) *no violence at all* to (4) *very violent*. In addition, they indicated how frequently they played each game, using a scale from (1) *almost never* to (5) *almost every day*. The scales showed adequate internal consistency ($\alpha = .61$).

Second, expert ratings were obtained for the listed games. For each game, a video clip of game play was created or found, and trained raters rated eight categories of violence for each game: Aggressive acts by the player, aggressive acts directed toward the player, rate of violent acts per minute, humanoid targets, blood and gore, use of weapons to harm, photorealistic violence, body parts severed, torn, or exposed. The presence or absence of each category was noted and a sum score was created. Because the frequency was multiplied by the violence score, it was necessary to avoid the value zero on this scale. Therefore, a value of one was added to every score yielding a range of (1) to (9). A single expert coder rated all of the 1,382 games for which video clips could be found, and a second rater coded 61% of the clips. Based on the high level of agreement between the raters (see Table 1), the first coder's ratings were used for all further analyses. Third, official ESRB ratings in the U.S. were gathered for each nominated game.

To examine the convergent and discriminant validity of the measures of violent game play, scores of violent and nonviolent game content were generated from each of the three measures. For the user ratings, all games that had received violence ratings of (2) or more on the scale from (1) to (4) were included in the category of violent games, all games with user ratings of violent content of equal to (1) were classified as nonviolent. For each game, the product of user ratings of violent content and their reported frequency of use was computed, and scores of violent and nonviolent game play were formed by averaging the product scores across the violent and nonviolent games, respectively.

A parallel categorization was used for the expert ratings. Games that had received an expert rating of violent content of less than (2) on the scale from (1) to (9) were categorized as

[1] Findings from these datasets addressing issues other than those at the core of this paper have been reported in Gentile et al. (2009) for the U.S.; Gentile, Choo et. al. (2011) for Singapore; and Krahé, Busching, and Möller (2012) for Germany.

nonviolent, games with expert ratings of (2) or above were included in the violent games category. For each game, the expert violence rating was multiplied by user frequency reports, and violent and nonviolent game play scores were created by averaging the product scores across the violent and nonviolent games, respectively. For the ESRB ratings, all games rated as *early childhood* (EC) or *everyone* (E) were classified as nonviolent, all other games were included in the category of violent games because over 90% of games rated E10+, T, or M include violent content (Gentile, 2008).

Self-reports of physical aggression. Children were asked whether they had been involved in a physical fight in the previous year (no (0) or yes (1)). This approach has been used successfully with children in other studies (Gentile, Coyne, & Walsh, 2011; Gentile et al., 2004).

Teacher-reports of aggression. Teachers completed a survey assessing the frequency of each child's observed aggression (Anderson et al., 2007). Teachers responded to four items that measured aggression on a 5-point scale, with responses ranging from (1) *never true* to (5) *almost always true*: This child: (a) hits or kicks peers, (b) initiates or gets into fights with peers, (c) pushes, or (d) shoves peers; threatens to hit or beat up other children. Internal consistencies were good at both Time 1 ($\alpha = .93$) and Time 2 ($\alpha = .93$).

Singapore

Sample. Children were recruited from 12 public schools in Singapore from diverse areas of the country. Informed consent was obtained from the parents through the schools. A liaison teacher from each school collated the information and excluded students from the study whose parents had refused consent. Assent was obtained from the students through informing them that participation in the survey was voluntary and they could withdraw at any time. Privacy of the students' responses was assured by requiring the teachers to seal collected questionnaires in the envelopes provided in the presence of the students.

Participants were 3,034 children and adolescents who were in 4rd, 5th, 8th, or 9th grades at the first wave of data collection. Seventy-two percent of the participants were male (2,193 male, 804 female, and 37 unknown). The mean age was 12.2 years ($SD = 2.1$, range 8 to 17 years of age) at the first wave of data collection. Participation rate was 99%. Children completed questionnaire measures in school twice 12 months apart. Questionnaire data were available for 3,034 and 2,360 participants for Time 1 and Time 2, respectively.

Measures.

Exposure to violence in games. At each wave, participants listed the three video games that they played the most. For each game, they indicated (a) how often they killed creatures and (b) how often they killed players in the game, using a four point scale ranging from (1) *never* to (4) *almost always*. Violent video game exposure was calculated as the mean frequency of killing creatures and players across the three games and also showed adequate internal reliability ($\alpha = .77$). In addition to self-reports of violent content, official ESRB ratings from the U.S. were gathered for each nominated game.

Measures of violent and nonviolent game play were derived in the same way as in the U.S. sample. Only games for which the users reported never shooting a creature or another player were considered nonviolent. Again, measures of violent and nonviolent game play were calculated by computing the product of user ratings of violence and frequency of play across the games in the violent and nonviolent categories, respectively. For the ESRB ratings, games rated suitable for *early childhood* (EC) or for *everyone* (E) were considered nonviolent, the remaining games were classified as violent.

Self-reported physical aggression. A modified version of the Social Interaction Survey (Crick, 1995) was used to measure physical aggression at both Time 1 and Time 2 (T1 $\alpha = .85$, T2 $\alpha = .86$). This 6-item scale asks participants to rate how true the statements are of them in the past year using a scale from (1) *not at all true* to (7) *very true*. An example item was "When someone makes me really angry, I push or shove that person." The items were summed, such that higher scores indicate higher levels of physical aggression.

Normative beliefs about aggression. A 20-item scale from Huesmann and Guerra (1997) was used to measure participants' normative acceptance of aggressive behavior under different types of provocations. Each item was rated on a 4-point scale, ranging from (1) *it's really*

wrong to (4) *it's perfectly OK* for scenarios such as "Suppose a boy says something bad to another boy, John, do you think it's wrong for John to hit him?" Internal consistency was high at both Time 1 and Time 2 (αs = .95).

Germany

Sample. Participants were recruited between January and April 2008 from 14 schools in different districts of Berlin, Germany. Approval for the study and all materials was obtained from the Ethics Committee of the University of Potsdam as well as from the school regulating body in Berlin. Active consent was obtained from all participants and additionally from parents of the students under the age of 14, in compliance with the general consent regulations. Response rates were very high, with no more than one or two students per class at most not participating.

In total, 1,715 secondary school students (881 female and 834 male) from 93 classes who were in 7th grade (N = 839) or 8th grade (N = 876) at Time 1 took part in a multiwave longitudinal study. The first two waves are included in the present analysis. The mean age of the sample was 13.4 years (SD = .87) at the first wave of data collection. Seventy-two percent of the participants took part in both waves. The majority of participants were German nationals (75.1%), 10.4% were Turkish nationals, 7.0% had a dual citizenship, and the remaining participants came from a range of different countries.

Measures.

Exposure to violence in games: Genre measure. Participants were provided with a genre list for video games derived from Möller and Krahé (2009). For each item on the list, they were asked to indicate how frequently they used the respective genre on a 5-point scale ranging from (0) *never* to (4) *very often*. Eleven genres were presented (see below), and each genre was illustrated by a specific title prominent at the time when the data were collected. Independent ratings of the violence level of each genre were obtained from media experts who rated each genre on a 5-point scale from (1) *nonviolent* to (5) *very violent*. Experts received the following definition of violent content: "By 'violent media depictions' we mean the intentional harming of humans, humanoid characters or other beings by one or more media characters/players. Based on this definition, a genre contains a lot of violence if it typically presents battle scenes or fights where characters hit, shoot at, injure, and/or kill others, where there is plenty of blood, and where scenes of injuring and killing others are presented in a realistic way." The experts were three sales persons, one woman and two men working for retailers specializing in video games. All three indicated that they played video games a lot both at home and in the context of their job, with an average of 13.7 years of gaming experience.

To arrive at a measure of *genre-based violent game play*, the seven video game genres that contained some measure of violence as reflected in expert violence ratings of higher than (2) on the 5-point scale from (1) *nonviolent* to (5) *very violent* (action adventure, military strategy, genre mix [a combination of shooter and racing games such as *Grand Theft Auto*], beat-em ups, role playing games, shooters, and survival horror games) were multiplied by participants' reported frequency of use. The resulting product scores were then averaged across the genres to yield a total measure of *genre-based violent game play.* The four game genres with expert violence ratings of less or equal two (construction strategy, classic adventure, simulations, and sports games) were combined into an overall index of *nonviolent game use*. The αs for violent video game use were substantial at both times, α = .86 for Time 1 and α = .88 for Time 2, suggesting that preferences for violent media contents show a consistent pattern across genres. The internal consistency for nonviolent game use was lower with α = .65 at Time 1 and α = .70 at Time 2, which was to be expected because of the greater heterogeneity of this measure across genres.

Exposure to violence in games: Free nominations. Participants were also asked to name their two favorite video games and rate them in terms of violent content: "How much violence does this game contain (e.g., how often do characters fight, how much blood and gore is shown)?" Responses were made on a four-point scale ranging from (1) *nonviolent* to (4) *very violent.* Expert violence ratings for the listed titles were obtained independently from a group of media experts (10 men and 2 women) who were either sales persons, working for retailers specializing in video games, or journalists working for

video game magazines. All indicated that they played video games a lot both at home and in the context of their job, with an average of 10.9 years of gaming experience. At least three expert ratings were obtained for each title. A mean violence score for each title was calculated across all experts. In addition, each title was assigned to a genre as defined in the USK database. For the USK genres for which a matching category was contained in our genre list (91% of all listed games), the respective expert violence rating for our genres was coded as an additional index of violent content. Unlike in the data sets from the U.S. and Singapore, the German study did not collect frequency ratings of using the listed games from participants.

Self-reported physical aggression. Self-reports of physical aggression were obtained at Time 1 and Time 2. Students indicated on a five-point scale from (0) *never* to (4) *very often* how often *in the past six months* they had (a) pushed, (b) kicked, (c) hit another person, (d) pulled another person's hair or scratched or bitten him or her, and (e) broken things on purpose that belonged to another person (see Krahé & Möller, 2010; $\alpha = .81$ for Time 1 and .82 for Time 2).

Teacher reports of aggressive behavior. For each participant, class teachers provided a rating of aggressive behavior on a single item: "How often does this student behave in an aggressive way toward others?" Ratings were made on a five-point scale ranging from (0) *never* to (4) *very often*. Teachers were asked to make these rating based on the student's behavior in the last school term (i.e., the last 6 months).

Normative beliefs about aggression. Beliefs about the normative acceptance of aggression were measured using a vignette that described a provocation scenario based on Krahé et al. (2012) and provided five possible reactions. Participants were asked to indicate how acceptable it would be for them to respond in that particular way in the situation. Two items represented physical aggression (e.g., "to kick and push him or her"), and three responses reflected relational aggression (e.g., "to spread rumors about him or her"). Responses were made on a 4-point scale ranging from (0) *not at all ok* to (3) *totally OK*. Reliability was good at both times with $\alpha = .80$ for Time 1 and $\alpha = .72$ for Time 2.

Results

Reliability

Ratings of violent content in video games can be said to be reliable to the extent that different raters arrive at similar assessments of the violence level of a game. Therefore, intercoder agreement was analyzed as an index of reliability, using the intraclass correlation coefficient.

Table 1
Interrater Agreement About Level of Violent Content in Video Games (Intraclass Correlations)

Measure	ICC	Question
U.S.		
Game violence-user rating	.61	How violent is this video game?
Game violence–expert rating	.84	Are the following things present in the game: Aggressive acts be the player, humanoid targets, blood and gore, etc.?
Singapore		
Game harm to creatures-user rating	.35	How often do you shoot or kill creatures in this game?
Game harm to players-user rating	.45	How often do you shoot or kill other players in this game?
Mean harm to creatures and players-user rating	.42	Mean of the above two items
Germany		
Game violence-user rating	.71	How much violence does this game contain (e.g., how often do characters fight, how much blood and gore is shown)?
Game violence-expert rating	.58	How much violence does this game contain?
Violent content of game genre-expert rating	.81	How much violence do games of this genre typically contain?

More specifically, interrater agreement was calculated for three measures of violent content: (1) genre ratings by experts, (2) expert ratings of individual game titles, and (3) user ratings of individual game titles. The results of the analysis are presented in Table 1.

In the U.S. and Germany, the ICCs were in the range of .60 to .80, which indicates moderate to strong agreement (LeBreton & Senter, 2008). The user ratings from Singapore showed lower ICCs, which ranged between .35 and .45. This is probably the result of the specific wording of the questions. Whereas the U.S. and German measures asked for global ratings of violent content, the Singapore measure asked more specifically about each individual's approach to game play. Because there is variation in game play, it is not surprising that this measure of violence showed a higher variance between different raters, resulting in lower interrater agreement. Overall, however, the data reflect a good consensus within each group of raters about the level of violent game content.

Validity

Convergent validity. Intercorrelations between the different operationalizations of violent video game use were assessed to determine the convergence of the measures in the three data sets. Table 2 presents the correlations for each country.

These findings demonstrate a high level of agreement within each country. By comparing the correlation coefficients, it is possible to see whether the different groups of raters share a common understanding of violent content or whether there are systematic differences between them. In the data from the U.S., the coefficients ranged between $r = .47$ and $r = .60$, which indicates a fairly high agreement. The data from Singapore showed a different picture. The two very high correlations ($r = .84$ and .86) of the scale mean with the respective items are not surprising, because the mean is a composite of the two items. The ESRB ratings showed a low correlation with the user ratings of violence against creatures ($r = .28$), but a somewhat higher correlation of $r = .43$ with ratings of violence against players and with the composite score. This suggests that the ESRB ratings mainly focus on violence between players and to a lesser extent on violence against game creatures. The German data showed a very high agreement between the different raters (*r*s from .79 to .91) without any outliers. Thus, there is clear evidence that the different measures of violence converge within each country.

In addition, convergent validity was examined across the three data sets for 39 game titles that had each been listed in the free user nominations in all three countries. By comparing ratings of violence levels for the same games from the U.S., Singapore, and Germany, it was possible to identify cultural similarities and differences in the perception of game violence

Table 2
Descriptive Statistics of Violence Ratings for Individual Titles From Different Sources

	Range	*M* (*SD*)	1	2	3
U.S.					
1. Game violence—user rating	1–4	1.81 (1.00)			
2. Game violence—expert rating	1–9	4.20 (2.48)	.57		
3. Game—ESRB rating	1–5	2.67 (1.04)	.60	.46	
Singapore					
1. Game harm to creatures—user rating	1–4	2.44 (1.32)			
2. Game harm to players—user rating	1–4	2.08 (1.25)	.45		
3. Mean harm to creatures and players—user rating	1–4	2.27 (1.10)	.86	.84	
4. Game—ESRB rating	1–6	3.24 (1.17)	.28	.43	.43
Germany					
1. Game violence—user rating	1–4	2.20 (1.15)			
2. Violent content of game genre—expert rating	1–5	2.52 (1.38)	.81		
3. Game violence—expert rating	1–5	2.47 (1.42)	.80	.90	
4. Game—USK rating	1–5	2.66 (1.45)	.79	.89	.91

Note. All correlations $p < .001$.

unaffected by differences in game content. The correlations are presented in Table 3.

The findings demonstrate a high level of agreement within each country and, importantly, a high cross-cultural consensus in the perception of the violence contained in the different games, as reflected in significant correlations ranging from $r = .59$ to $r = .90$. This is evidence of a common cross-cultural understanding of media violence as reflected in popular video games.

Construct validity or predictive validity. To establish construct validity, we examined how similar the different operationalizations of violent video game play were in predicting aggression-related outcome measures. These analyses examined both the cross-sectional and the longitudinal correlations of the different measures of violent video game use with various aggression-related outcome variables, controlling for nonviolent game play. The more similar the measures are in their links with aggression, the greater confidence we can have in their validity in representing the underlying construct of exposure to violence in video games. Only those measures for which both violence ratings and frequency ratings were available could be used for this analysis as the level of violent content alone, regardless of frequency of use, is not expected to be strongly linked to aggression. Because the German data set contained user frequency ratings only for the genre lists, not for the individual titles, only the genre-based index of violent and nonviolent game play was used from the German data set for these analyses.

To examine the links between the different operationalizations of violent video game play and aggressive behavior as well as normative beliefs about aggression, partial correlations were calculated, controlling for nonviolent video game use. The partial correlations are displayed in Table 4. The majority of correlations ranged between $r = .15$ and $r = .25$. More important than the absolute values of the correlations is the finding that they were of similar magnitude across the measures used, indicating a satisfactory agreement between the different approaches.

However, there were systematic differences as well. For the U.S. and Singapore samples, the user rated measure of violent play yielded consistently larger cross-sectional and longitudinal effects than the ESRB ratings. Of special interest is the fact that within the U.S. sample, this occurred with both the self-reported measure of aggression and the teacher-reported measure of aggression. This suggests that the user rating measure of violent game play is not artificially correlated with aggression because of common method of measurement factors. One explanation of the smaller effects obtained by the ESRB method may be that ESRB ratings consider multiple types of content in addition to violent content.

In the comparison across countries, it is noteworthy that the highest correlations with the aggression-related outcomes were found for the genre-based measure from the German data. Because the same pattern can be found for the relations between violent game play and normative beliefs as well as teacher-rated aggression at Time 2, this cannot be attributed to common method bias.

Table 3
Correlations of Violence Ratings for Individual Games Within and Across Cultures

	1	2	3	4	5	6
1. U.S.—user rating						
2. Singapore—user rating (mean harm)	.69					
3. Germany—user rating	.80	.69				
4. U.S.—expert rating	.76	.76	.78			
5. Germany—expert rating	.80	.69	.87	.82		
6. U.S.—ESRB rating	.81	.59	.73	.68	.77	
7. Germany—USK rating	.82	.69	.86	.80	.90	.81

Note. The figures in the boxes indicate agreement within the same data source across countries. Listwise deletion, $N = 39$. All correlations significant at $p < .001$.

Table 4
Partial Correlations Demonstrating Predictive and Discriminant Validity

	Cross-sectional associations			Longitudinal associations		
	T1 Normative beliefs about aggression	T1 Self-reported physical aggression	T1 Teacher-reported aggression	T2 Normative beliefs about aggression	T2 Self-reported physical aggression	T2 Teacher-reported aggression
U.S.						
Violent play-user rating		.17***	.26***		.18***	.23***
Violent play-expert rating		.06	.19***		.15***	.14***
Violent play-ESRB rating		.12***	.20***		.13***	.19***
Nonviolent play-user rating		.10*	.04		.04	.08
Nonviolent play-expert rating		.03	.12**		.06	.03
Nonviolent play-ESRB rating		.08*	.01		.04	.05
Singapore						
Violent play-user rating of violence against creatures	.11***	.18***		.12***	.13***	
Violent play-user rating of violence against players	.15***	.21***		.16***	.15***	
Violent play-mean of user rating	.13***	.21***		.15***	.15***	
Violent play-ESRB rating	.08**	.09***		.06**	.04	
Nonviolent play-user rating of violence against creatures	−.03	.01		−0.02	.00	
Nonviolent play-user rating of violence against players	−.04	−.01		−0.03	.00	
Nonviolent Play-mean of user rating	−.05*	−.01		−0.03	.00	
Nonviolent Play-ESRB rating	.02	.02		0.03	.02	
Germany						
Violent play-genres (list measure)	.29***	.38***	.21***	.23***	.34***	.27***
Nonviolent play-genres (list measure)	−.05	−.08*	.00	−.03	−.07*	−.06

Note. All correlations between violent and nonviolent play are corrected for nonviolent play, and all correlations for nonviolent play are corrected for violent play.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Given that the impact of violent game content is assumed to increase with dosage, the information on how frequently players use violent video games is critical for the assessment of total media violence exposure. Therefore, in each of the measures reported in Table 4, frequency of use was taken into account. If only the level of violent content was considered regardless of frequency of exposure, the correlations between the different measures of violent game use and the aggression-related outcomes decreased in the U.S. data set on average by .02 and in the Singapore data set by .05. While the absolute decrease seems small, it is found consistently across almost all reported correlations and since most relationships are in the small to medium range, even small increases in predictive validity are important.

Discriminant validity. A final aim was to examine whether measures of violent video game play would differ from measures of nonviolent video game play as well as measures of overall gaming in their links with the selected aggression-related outcome variables. If, as claimed for instance by the GAM, the link between video game use and aggression is primarily driven by the violent content of the games, weaker associations should be found between measures of nonviolent video game play and aggression. To test this proposition, the partial correlations between nonviolent game use and the same aggression-related outcomes were calculated, controlling for violent media use. These results are also displayed in Table 4. With just three exceptions, all correlations of nonviolent game play and the aggression measures were nonsignificant, and the three significant correlations were substantially smaller than the corresponding correlations for violent games. These findings demonstrate the discriminant validity of the measures of violent video game play examined in our study.

Discussion

The primary goal of this study was to test the reliability, validity, and convergence of several different methodological approaches to measuring violent video game exposure, using longitudinal data from three cultures (U.S., Singapore, and Germany). Several measurement approaches have been used, including user ratings of violent content, official agency ratings, and expert ratings of specific games or genres. Most of the published studies use only one type of measurement. This renders comparisons between studies potentially difficult because it remains unclear to what extent differences in the links with aggression are the result of differences in the operationalization of violent game play. Our study had the strength of using multiple methods both within and across cultures. Overall, we found that almost all approaches to measurement showed sufficiently high reliability, convergent validity, predictive validity, and discriminant validity.

At first it may appear surprising that the results are so stable and consistent, given the differences between cultures, ages of participants, and measurement approaches, especially given the controversy that appears to surround violent video games. Quality of measurement is one feature that has been argued to have an impact on the results, with some researchers claiming that poorer measurement artificially diminishes the relations between violent game play and aggressive outcomes (e.g., Anderson et al., 2010) and others claiming that it artificially enhances these relations (e.g., Ferguson, 2010). It appears that this actually may not be a particularly important confound, at least where measurement of violence exposure is concerned. The caveat, however, is that some studies only measure the total amount of play and do not measure violent content directly, do not measure frequency or amount of time playing violent games, or that they do not control for nonviolent game play. As demonstrated here (and consistent with theoretical predictions), such measurement shortcomings *will* artificially lower the relations between violent game play and aggression outcomes. Furthermore, using ESRB ratings also consistently resulted in smaller effects than relying on user ratings, suggesting that studies using the ESRB approach may systematically underestimate actual effect sizes.

These results seem to be in accord with the majority of empirical results, including many of the studies that demonstrate smaller effect sizes. Some of those studies measure total amount of gaming or rely on the ESRB ratings, which consider criteria other than violence. Furthermore, all of the meta-analyses (including those of the critics of the literature) find about the same effect sizes empirically. Some studies

have compared Eastern and Western cultures, examining whether there may be differences in violent game effects because base rates of aggressive behavior vary widely between cultures. However, despite differences in base rates, the effects seem to be similar across cultures (Anderson et al., 2008, 2010). The present study provides additional evidence for the stability across cultures and, by implication, for the generalizability of violent video game research.

Best Practice Recommendations

Based on the present findings, eight best practice recommendations are suggested for the future measurement of violent content in video game research. Some of the recommendations may appear obvious, but they are by no means trivial given their disregard in published research (see Anderson's, 2004, and Anderson et al. 2010 meta-analyses distinguishing between best-practice and not-best practice studies).

(1) First, it is critical that researchers seeking to test violent video game effects measure violent content rather than simply measuring total amount of play or including games with little or no violent content.

(2) Second, because the amount of play and the content of play are not conceptually isomorphic and empirical studies demonstrate that they can have very different effects (Gentile, 2011), amount of play should be measured in addition to game content, for example by obtaining frequency ratings of using a particular game or genre (as suggested by the meta-analysis by Anderson et al., 2010). Users can play moderately violent games regularly, or they may play very violent games every now and then. The resulting exposure to violence in the virtual reality of video games would be similar in both cases. Therefore, violence exposure should always include both aspects, and it seems sensible to multiply these two aspects with each other.

(3) Third, it is advisable that researchers measure nonviolent play and control for it to focus the analyses on violent content. This could be achieved, as demonstrated here, by creating both violent and nonviolent exposure scores.

(4) Fourth, the findings show that both asking participants to list their favorite game titles and using genre lists for which violence ratings are available provides reliable and valid measures of violent game play. Therefore, researchers can choose between these approaches.

(5) Fifth, they can also choose between relying on user ratings or expert ratings of violent content in video games, as both sources provide reliable and valid characterizations of the level of violence in the video game diet of research participants.

(6) Sixth, when user ratings for specific games are obtained, it is preferable to use measures that ask about the general properties of the game (i.e., "how much violence is there in this game?") rather than the specific violent actions performed by the player ("how often do you shoot or kill . . ."), as these refer more closely to the properties of the game. The present results show a lower reliability for questions asking about individual game play, which are more susceptible to individual differences as well as desensitization because of prolonged exposure to violent media. However, the impact of this increase in error variance seems small, since the pattern of correlations with the aggression-related outcomes is very similar. Therefore, if the question of interest focuses on individual differences of game play, the latter may be preferable.

(7) Seventh, official game ratings appear to be much more culture-dependent than user ratings or expert ratings, as reflected in the low correlations of the U.S. ERSB ratings with user ratings in the dataset from Singapore. Therefore, they should be used only in the country in which they were developed. However, the consistently lower predictive validity of the ESRB method in the U.S. sample (relative to user ratings), coupled with the confounding of sexual and violent content in the ESRB rating scheme, suggest that the user rating method may be preferred. A general qualification regarding the reliance on official agency ratings refers to the fact that our analysis did not examine whether these ratings provide appropriate age recommendations for practical purposes. We tested only whether the rank order implied in the age ratings corresponded to the rank order based on expert or user ratings of the same titles. Whether the age recommendations are appropriate from a psychological point of view is beyond the scope of our study.

(8) Researchers also need to consider the time and expense of the different methods of measuring exposure to violent video games. In many re-

search contexts, the simplest method is to obtain user ratings of violent content, either of specific games or genres. The present results suggest that despite the expected inconsistencies of game ratings across participants, this approach is valid, and it does not require the extensive coding that the expert ratings or established rating systems require. When user time (or number of items) is a major issue, then it is reasonable to use expert ratings of violent content, even though doing so may be more expensive and require more time by the research team.

We believe that this research has several strengths: It includes large samples of children and adolescents from North America, Europe, and Asia. Furthermore, each sample used several measures of violent video game content in the same study, allowing for direct comparisons both within and between samples. Finally, each sample was longitudinal with at least a 12-months lag between assessments to examine the predictive validity of different measures of violent video game use for aggression-related outcome variables. Despite these strengths, however, this study is limited by its correlational nature. Participants were not randomly assigned to playing either violent or nonviolent games, but simply reported on their in vivo game play. Strong causal claims cannot be made from this research. However, the primary purpose of this study was not to test the effects of violent games on aggression-related variables, but to compare the different operationalizations of violent game exposure, thereby helping to clarify apparent discrepancies in the literature.

In summary, the construct of violent video game exposure is robust to measurement approaches across multiple cultures and age groups. At the same time, our analysis has provided several avenues in which future research may be improved. We hope that the empirically derived best practice criteria proposed on the basis of this analysis will be of benefit to future research.

References

Anderson, C. A. (2004). An update on the effects of playing violent video games. *Journal of Adolescence, 27,* 113–122. doi:10.1016/j.adolescence.2003.10.009

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science, 12,* 353–359. doi:10.1111/1467-9280.00366

Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology, 53,* 27–51. doi:10.1146/annurev.psych.53.100901.135231

Anderson, C. A., & Dill, K. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78,* 772–790. doi:10.1037/0022-3514.78.4.772

Anderson, C. A., Gentile, D. A., & Buckley, K. E. (2007). *Violent video game effects on children and adolescents: Theory, research, and public policy.* New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195309836.001.0001

Anderson, C. A., Sakamoto, A., Gentile, D. A., Ihori, N., Shibuya, A., Yukawa, S., . . . Kobayashi, K. (2008). Longitudinal effects of violent video games on aggression in Japan and the United States. *Pediatrics, 122,* e1067–e1072. doi:10.1542/peds.2008-1425

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H., . . . Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin, 136,* 151–173. doi:10.1037/a0018251

Crick, N. R. (1995). Relational aggression: The role of intent attributions, feelings of distress, and provocation type. *Development and Psychopathology, 7,* 313–322. doi:10.1017/S0954579400006520

Ferguson, C. J. (2007a). Evidence for publication bias in video game violence effects literature. *Aggression and Violent Behavior, 12,* 470–482. doi:10.1016/j.avb.2007.01.001

Ferguson, C. J. (2007b). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly, 78,* 309–316. doi:10.1007/s11126-007-9056-9

Ferguson, C. J. (2010). Blazing angels or resident evil? Can violent video games be a force for good? *Review of General Psychology, 14,* 68–81. doi:10.1037/a0018941

Ferguson, C. J. (2011). Video games and youth violence: A prospective analysis in adolescents. *Journal of Youth and Adolescence, 40,* 377–391. doi:10.1007/s10964-010-9610-x

Funk, J. B., Buchman, D. D., Jenks, J., & Bechtoldt, H. (2003). Playing violent video games, desensitization, and moral evaluation in children. *Journal of Applied Developmental Psychology, 24,* 413–436. doi:10.1016/S0193-3973(03)00073-X

Gentile, D. A. (2008). The rating systems for media

products. In S. Calvert & B. Wilson (Eds.), *Handbook of children, media, and development* (pp. 527–551). Oxford, England: Blackwell Publishing. doi:10.1002/9781444302752.ch23

Gentile, D. A. (2011). The multiple dimensions of video game effects. *Child Development Perspectives, 5,* 75–81. doi:10.1111/j.1750-8606.2011.00159.x

Gentile, D. A., Choo, H., Liau, A., Sim, T., Li, D. D., Fung, D., & Khoo, A. (2011). Pathological video game use among youths: A two-year longitudinal study. *Pediatrics, 127,* e319–e329. doi:10.1542/peds.2010-1353

Gentile, D. A., Coyne, S. M., & Walsh, D. A. (2011). Media violence, physical aggression and relational aggression in school age children: A short-term longitudinal study. *Aggressive Behavior, 37,* 193–206. doi:10.1002/ab.20380

Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of Adolescence, 27,* 5–22. doi:10.1016/j.adolescence.2003.10.002

Gentile, D. A., Maier, J. A., Hasson, M. R., & de Bonetti, B. L. (2011). Parents' evaluation of media ratings a decade after the television ratings were introduced. *Pediatrics, 128,* 36–44. doi:10.1542/peds.2010-3026

Gentile, D. A., Welk, G., Eisenmann, J. C., Reimer, R. A., Walsh, D. A., Russell, D. W., . . . Fritz, K. (2009). Evaluation of a multiple ecological level child obesity prevention program: Switch what you do, view, and chew. *BMC Medicine, 7,* 49. doi:10.1186/1741-7015-7-49

Höynck, T., Mössle, T., Kleimann, M., Pfeiffer, C., & Rehbein, F. (2007): *Jugendmedienschutz bei gewalthaltigen Computerspielen: Eine Analyse der USK-Alterseinstufungen*. [Youth media protection for violent computer games: An analysis of USK age ratings] KFN-Forschungsbericht; Nr.: 101. Hannover: Kriminologisches Forschungsinstitut Niedersachsen.

Huesmann, L. R., & Guerra, N. G. (1997). Children's normative beliefs and the development of aggressive behavior. *Journal of Personality and Social Psychology, 72,* 408–419. doi:10.1037/0022-3514.72.2.408

Hyman, P. (2005). Rated and willing: Where game rating boards differ. *Game Developer magazine, 8*. Retrieved from http://www.gamasutra.com/view/feature/130896/rated_and_willing_where_game_.php

Krahé, B., Busching, R., & Möller, I. (2012). Media violence use and aggression among German adolescents: Associations and trajectories of change in a three-wave longitudinal study. *Psychology of Popular Media Culture, 1,* 152–166. doi:10.1037/a0028663

Krahé, B., & Möller, I. (2004). Playing violent electronic games, hostile attributional style and aggression-related norms in German adolescents. *Journal of Adolescence, 27,* 53–69. doi:10.1016/j.adolescence.2003.10.006

Krahé, B., & Möller, I. (2010). Longitudinal effects of media violence on aggression and empathy among German adolescents. *Journal of Applied Developmental Psychology, 31,* 401–409. doi:10.1016/j.appdev.2010.07.003

Kutner, L. A., & Olson, C. K. (2008). *Grand theft childhood: The surprising truth about violent video games and what parents can do*. New York, NY: Simon & Schuster.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. doi:10.1177/1094428106296642

Möller, I., & Krahe, B. (2009). Exposure to violent video games and aggression in German adolescents: A longitudinal analysis. *Aggressive Behavior, 35,* 75–89. doi:10.1002/ab.20290

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. doi:10.1037/0021-9010.88.5.879

Richmond, J., & Wilson, J. C. (2008). Are graphic media violence, aggression and moral disengagement related? *Psychiatry, Psychology and Law, 15,* 350–357. doi:10.1080/13218710802199716

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *The Observer, 9,* 27–29.

Sherry, J. L. (2001). The effects of violent video games on aggression. *Human Communication Research, 27,* 409–431. doi:10.1111/j.1468-2958.2001.tb00787.x

Wei, R. (2007). Effects of playing violent videogames on Chinese adolescents' pro-violence attitudes, attitudes toward others, and aggressive behavior. *CyberPsychology & Behavior, 10,* 371–380. doi:10.1089/cpb.2006.9942

Received February 20, 2013
Revision received May 5, 2013
Accepted May 23, 2013 ■