CHRISTOPHER J. FERGUSON

3. THE WILD WEST OF ASSESSMENT

Measuring Aggression and Violence
in Video Games

INTRODUCTION

The issue of violent video game and their potential influence on young and adult players has been a "hot" topic in the public arena (Thompson, 2007) as well as scientific discourse (Kutner & Olson, 2008). However as Ioannidis (2005) has noted, research fields that are new or "hot" are particularly prone to a variety of biases including publication bias, citation bias and poor methodology that can result in misleading conclusions that prove unreliable over time. Increasingly these issues are becoming apparent in video game violence studies wherein conclusive statements about causal links between video game violence exposure and acts of aggression or violence on the part of players may have greatly preceded evidence available to support such claims, which remains in short supply (Ferguson, 2008; Grimes, Anderson & Bergen, 2008; Olson, 2004). Among the issues identified as problematic in video game violence studies is the issue of assessment. As noted by Ferguson (2007), and at the risk of some hyperbole, assessments in the realm of video game violence studies have too often been shockingly and embarrassingly poor, with common use of unstandardized, and poorly validated measures, upon which unscientific claims of generalizability to real world acts of extreme violence are made. The current chapter reviews the issue of assessments in violent video game studies and offers suggestions for improvements in this realm.

DEFINING RELEVANT TERMS

Before discussing issues related to "aggression" and "violence" we must come to some kind of understanding about what these terms mean and how they should be defined. Aggression has been defined as behavior produced to cause physical harm or humiliation to another person who wishes to avoid it (Baron & Richardson, 1994). Although reasonable on the surface, this definition is limited by a rather obvious bias: namely the implication that aggression is always "bad". In other words, the definition above is defined in such a way as to imply that the aggressor is a "perpetrator" and the aggression recipient is a "victim." As such, this is an incomplete and fundamentally flawed definition of aggression. It is implied that aggression has no adaptive function and is always pathological and undesirable. This would appear to be naïve, and at best is an assumption. In moderate doses, aggression may very well be adaptive, guiding individuals toward many behaviors approved of by society including standing up for

one's beliefs, assertiveness, defending others in need, careers in law enforcement, the military, business, legal affairs, etc., sporting activities, political involvement, debate and discourse indeed including scientific debate (Hawley & Vaughn, 2003; Smith, 2007). For purposes of this discussion a slightly altered definition of aggression is proposed, namely that aggression is "Behavior which is intended to increase the social dominance of the organism relative to the dominance position of other organisms" (Ferguson & Beaver, 2009). Activities which met Baron and Richardson's (2004) definition of aggression would still fall within the current definition, although the current definition is stripped of moral implications. Aggression, then, is behavior intended to increase one's own dominance and, thus, reproductive success. Other organisms may or may not be harmed depending on the form or intensity of the aggressive behavior. Violent behavior certainly would be aggressive, but not all aggressive behaviors are violent or even necessarily negative from a cultural perspective.

The World Health Organization (2002, p. 5) has defined violence as:

> The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation.

By and large this appears to be a reasonable definition of violence. It is worth noting that not all violent acts are negative. Violent acts motivated by self-defense, or defense of one's family, social group or culture are generally deemed as acceptable. Such violent acts may also be adaptive due to the protection of kin (Queller & Strassman, 2002; Smith, 1964) or, in the case of cultural violence, the advancement of one's cultural group and prestige for oneself, with consequential improvement of potential mating options. Although violent behaviors tend to carry significant risk of injury, at times they may be adaptive with the risks of not being violent greater than for engaging in violent behavior.

By contrast use of the term extreme violence specifically refers to violent behavior for which the risks outweigh potential benefits (Ferguson & Beaver, 2009). Risks either of personal injury or to one's social esteem through disapproval, retaliation, or incarceration are of likelihoods greater than any anticipated benefits. Extreme violence then is rather synonymous with criminal violence. Yet because criminal codes vary from state to state, country to country and from one time to another, use of the term "criminal violence" may be too subjective to be truly meaningful.

Naturally "violence" tends to be difficult to study empirically owing to both practical and ethical limitations. As such, researchers often fall back on a rather broad category of aggression measures, although researchers must be cautious not to assume such measures correspond well to more serious violent behaviors.

VIOLENT VIDEO GAMES: A BRIEF STATE OF THE LITERATURE

In a broad sense a violent video game is any game in which players have the opportunity to inflict imaginary or fantasy physical harm upon another character in the game, whether another player character or a non-player character (that is, a character controlled by the computer). The term "violent video game", which itself may be

inappropriately emotionally evocative for scientific discussion, typically conjures up images of highly controversial games such as the *Grand Theft Auto* series (see Thompson, 2007). Thompson and Haninger (2001) have demonstrated that mild levels of violence are common in games that are rated as appropriate for young children by the Entertainment Software Ratings Board (ESRB). From this finding we can glean that even young children find mildly violent content to be highly appealing (Olson, 2010). Indeed as Olson (2010) notes, many children may use violent games to unwind from stress or connect socially with peers. Indeed, in early research video games such as *Space Invaders* (Silvern & Williamson, 1987), *Centipede* and *Zaxxon* (Anderson & Ford, 1987) were labeled as "violent video games" although arguably such video games would likely illicit little concern from most parents or policy makers. Today's games eclipse previous generation games in both graphics and immersion. This sophistication growth curve allows for a higher level of realism in today's games that may account for their increased scrutiny. In contrast, the graphically inferior and less complex games of previous generations that were once considered violent have since been increasingly recognized as "harmless".

Despite operational definitions such as the one provided above, delineating which games are "violent" is not always clear. For instance, actions in sports games which follow the rules of the sport and are not intended to cause harm such as blocking or tackling have generally not been considered "violent" either by standard rating systems such as the Entertainment Software Ratings Board (ESRB) or by many scholars, although there may be some reasonable disagreement on this score.

Prior to approximately the year 2000, most scholars acknowledged that results from video game violence studies provided inconsistent results (Cooper & Mackie, 1986; Dominick, 1984; Kirsh, 1998; van Schie, & Wiegman, 1997). Two factors appear to have changed the scholarship of video game violence studies in approximately 1999–2000. The first was the Columbine High Massacre in which Eric Harris and Dylan Klebold killed a teacher and a number of classmates before committing suicide; a crime that was "linked" to the video game *Doom* (although since most male youth play violent video games, making such a "link" may be less meaningful than seems at first). Second, several scholars began making increasingly certain statements about the links between violent games and aggression (e.g. Anderson & Dill, 2000; Anderson & Bushman, 2001; Bartholow, Bushman & Sestir, 2006; Gentile, Lynch, Linder, & Walsh, 2004; Huesmann, 2007) despite the objections of other scholars (Olson, 2004; Sherry, 2007) who have argued that the data remain weak, inconsistent and unsupportive of the causal view. Indeed, since 2000, many studies continue to find no link between violent game playing and subsequent aggression (Colwell & Kato, 2003; Durkin & Barber, 2002; Ferguson, Rueda, Cruz, et al., 2008; Ferguson & Rueda, in press; Ferguson, San Miguel & Hartley, 2009; Funk, Hagan, Schimming, et al., 2002; Unsworth, Devilly & Ward, 2007; Williams & Skoric, 2005), although other authors have engaged in *citation bias*, simply ignoring this field of studies while continuing to expound a high degree of certitude regarding video game violence and player aggression (e.g. Anderson, 2004). For instance in speaking about testimony given in the ESA, VSDA and IRMA v. Blagojevich, Madigan and Devinecase involving the regulation of violent video game sales to minors, the court noted that:

With regard to their conclusions, Dr. Goldstein and Dr. Williams noted that Dr. Anderson not only had failed to cite any peer-reviewed studies that had shown a definitive causal link between violent video game play and aggression, but had also ignored research that reached conflicting conclusions. Dr. Goldstein and Dr. Williams noted that several studies concluded that there was no relationship between these two variables. They also cited studies concluding that in certain instances, there was a *negative* relationship between violent video game play and aggressive thoughts and behavior (e.g., initial increases in aggression wore off [   ] individual was allowed to play violent video game for longer period)." (ESA, VSDA and IRMA v. Blagojevich, Madigan and Devine, 2005, pp. 14–15).

The court also expresses similar concerns about cherry-picking of data by politicians involved in enacting anti-game legislation, noting (ESA, VSDA and IRMA v. Blagojevich, Madigan and Devine, 2005, p. 16):

Finally, the Court is concerned that the legislative record does not indicate that the Illinois General Assembly considered any of the evidence that showed no relationship or a negative relationship between violent video game play and increases in aggressive thoughts and behavior. The legislative record included none of the articles cited by Dr. Goldstein or Dr. Williams. It included no data whatsoever that was critical of research finding a causal link between violent video game play and aggression. These omissions further undermine defendants' claim that the legislature made "reasonable inferences" from the scientific literature based on "substantial evidence."

Other scholars have noted this issue as well. For example Sherry (2007) in speaking about the rhetoric on video game violence expressed "Further, why do some researchers (e.g. Gentile & Anderson, 2003) continue to argue that video games are dangerous despite evidence to the contrary?" (p. 244).

This unfortunate state of affairs has led some scholars (Grimes, et al, 2008; Kutner & Olson, 2008) to conclude that video game violence studies and media effects studies more generally are hopelessly politicized, with statements of causal certitude essentially in direct inverse proportion to the quality and consistency of the data available.

MEASUREMENT AND ASSESSMENT IN VIDEO GAME VIOLENCE STUDIES

Naturally, for causal statements to be made about links between violent video games and player aggression (or other outcomes such as visiospatial cognition or educational performance) assessments of relevant constructs must be standardized, reliable and valid. It is generally recognized that these three issues proceed in linear fashion. In other words an assessment instrument must, by definition, be standardized before it can be considered reliable, and must have demonstrated reliability before it can be proven to have reasonable validity. Standardization denotes that the assessment instrument is used in exactly the same manner, regarding both administration and scoring, every time it is used, not only by a single researcher, but across all research

protocols. Reliability, or the consistency of an assessment instrument, can be demonstrated in several ways although internal consistency coefficient alpha is arguably most dependable for the majority of behavioral measures (Nunnally & Bernstein, 1994). Kappa reliability for behavioral ratings are also common, and tend to perform similarly for most analyses of reliability, although it is cautioned that two forms of reliability should not be considered to be truly interchangeable. Test-retest and other forms of reliability may also be commonly used and may be equally or more important than coefficient alpha in some circumstances. Generally speaking, reliabilities as low as .70 are typically considered tolerable although higher reliabilities are desirable.

Validity can be more difficult to demonstrate, often requiring multiple studies. Multiple forms of validity exist, and won't be reviewed here, although validating new measures against well-validated existing measures, or against real-world behaviors or outcomes are typical means of demonstrating validity. Although standards for validity are arguably less clear than for reliability, validity coefficients of .40 or better are considered to be sufficient to demonstrate good validity (Ferguson & Rueda, 2009).

Through the time of this writing attention to these basic psychometric principles has been in short supply. Rather than initially focusing on the development of standardized and well-validated outcome measures, researchers instead have too often relied on "ad hoc" or "on the fly" outcome measures. Ad hoc measures are those that have little prior history of use and are commonly developed solely for use in a particular study (although they may be adopted by future studies, often without providing any further evidence for validity). For example, a researcher wishing to study childhood aggression experimentally naturally cannot ethically or legally incite children to acts of serious violence. With that limitation in mind, the researcher is forced to invent, ad hoc, a behavioral outcome that vaguely resembles violent behavior without being violent behavior. Perhaps, in this case, the researcher gives the child participants a pillow to hold on to, then counts the number of times the child swings the pillow around or throws it, rationalizing that these are "aggressive" actions. It all sounds vaguely reasonable…except that there is no data to support that outcomes related to swinging a pillow around can be generalized to serious acts of youth violence.

Unfortunately such ad hoc measures are tolerated to a great degree in media violence research. For instance studies of media violence have included ad hoc outcome measures such as popping a balloon, participants rating the job performance of a graduate student who had just insulted them, filling in the missing letters of words (so that filling in KN_V_S as "knives" rather than "knaves" would be considered "aggressive"), or completing the missing endings of children's stories (see Ferguson, 2009 for a discussion). It is unfortunate enough that such measures are blithely generalized to serious violent behaviors (e.g. Anderson, 2004), but they are deficient from a psychometric standpoint in that they not only commonly lack sufficient validity data, but also present options for unstandardized use. Such measures rarely have well-validated, published, detailed instructions for administration and scoring and when they are adopted for use by different studies, may be used differently between studies.

35

One of the most commonly used, and poorly standardized instruments used in laboratory studies of aggression, is the modified Taylor Competitive Reaction Time Test (TCRTT). The original TCRTT (seldom used in video game studies) involved the research participant playing a consensual reaction-time game against an opponent. The participant would set electroshock levels for their opponent to receive when they "lost" trials of the reaction time game, and would receive shocks in return for losing. In reality, the game was a ruse, with no actual opponent, and the series of wins and losses standardized no matter how fast the participant might be. The TCRTT was later modified with noise blasts (similar to white noise on a television set) replacing the electroshocks. The noise blasts are non-painful, potentially eliminating some ethical concerns, and without the need for an electroshock apparatus, could be administered easily on a common computer. The TCRTT involved 25 trials, in which the intensity and duration of the noise blast could be set independently before each trial.

Although varying scholars have tinkered with the administration of the TCRTT (Anderson & Dill, 2000; Konijn, Bijuank, Bushman et al., 2007) a more pressing concern is the absence of a standardized assessment protocol for the TCRTT. In fact, with a little creativity, a multitude of possible outcomes could be calculated. For instance one could average the intensity and/or duration across all 25 trials. Or one could impose an arbitrary cut-off (say 8 out of 10 on the intensity or duration meters) and count the trials that exceed that number. Or one could count only those scores occurring after "win" or "loss" trials. Or one could multiply (or otherwise manipulate) the two measures (intensity and duration) together in some fashion. Or one could examine only the first trial, and ignore the other 24, arguing the first trial is "unprovoked" aggression (since only on the first trial is the participant unaware how high their opponent is setting noise levels). In other words researchers have a multitude of potential "outcomes" from which to choose.

Although it may be controversial to say, I suggest it is not only possible but probable that researchers are able to select from among these those which best support their a priori hypotheses, particularly in such a highly politicized research field (see Ioannidis, 2005).

One need only to examine the research literature with the TCRTT to see that this is the case. Different researchers…and even sometimes the same researcher!...have used the TCRTT differently across studies (see Anderson and Dill 2000; Anderson & Murphy 2003; Bartholow et al. 2006; Carnagey and Anderson 2005 for four different ways of using the modified TCRTT). Several meta-analytic reviews have now found that studies that use measures such as the TCRTT in an unstandardized fashion produce higher effect sizes, than studies using standardized assessment measures (Ferguson, 2007; Ferguson & Kilburn, 2009). A standardized version of the modified TCRTT has been proposed (Ferguson et al., 2008) using only the intensity measure averaged across all 25 trials (the duration measure was found to have particularly poor validity), but whether the field will shift to a standardized approach remains to be seen.

If one ignores the issue of standardization (which this research field seems intent on doing!), we move to concerns about the validity of outcome measures. The field of video game violence studies is intended to generalize to real-world

clinically significant acts of serious violence. Any claims to the contrary are simply false, particular when mass school-shooting incidents are routinely invoked in manuscripts written by scholars in this research field (e.g. Anderson, 2004; Anderson & Dill, 2000; Barlett & Anderson, 2009; Grossman, 1996). For such lofty comparisons… between filling in the missing letters of words, or delivering non-painful noise bursts to an ostensibly consenting opponent to serious acts of youth violence or even mass homicides…validity data must be of the first caliber. Unfortunately the use of poorly validated outcome measures has been found to be a common problem (Ferguson & Kilburn, 2009; Kutner & Olson, 2008).

That aggression measures, particularly those used in laboratory designs, have struggled with severe validity problems has been known for some time (Freedman, 1992; Ritter & Eslea, 2005; Tedeschi & Quigley, 1996, 2000; Savage, 2004), nonetheless the field has failed to either resolve the validity problems or to restrain attempts to generalize poor measures to real-world behaviors. In fairness, developing the "better mousetrap" of aggression measures is likely to remain difficult, given the practical and ethical limitations of studying aggression in laboratory settings. Nonetheless, generalizations between poorly validated outcome measures and serious acts of real-world violence have approached the level of pseudoscience (Ferguson, 2008). Specific measures such as the modified TCRTT have been examined in regards to their ability to predict violent behaviors and have performed poorly (Ferguson & Rueda, 2009).

Validity problems may be less severe for correlational measures, which include self-reported aggression measures, parent, teacher or peer-report measures, clinical violence prediction measures (which in fact are never to seldom used in video game violence studies despite arguably being amongst the best outcome measures) or arrest record data. Even here scholars should attend to differences between ad-hoc surveys (those invented for the specific project and lacking validity data) compared to well-established clinical measures of aggression (such as relevant subscales on the Child Behavior Checklist or Personality Inventory for Youth). Survey instruments suffer significant limitations themselves. Self and parent report inventories may be subject to social desirability effects, where as peer nominations of aggression may devolve into "popularity contests" rather than valid aggression tools. Arrest record data similarly only provides a measure of crimes for which participants have been caught which likely underrepresents the total crimes an individual has engaged in.

Earlier in this chapter it was noted that aggression is not universally bad. In fact, it is likely that the adaptive function of aggression takes something of a curvilinear form, with individuals exhibiting moderate aggression actually most adaptive (Ferguson & Beaver, 2009). Thus, showing small increases or decreases in aggressive behavior, even if measured validly, are insufficient to document harm. Indeed harm is documented onto when elevations in aggression exceed clinically significant levels, that is levels likely to reduce the adaptive ability of the individual. In order to document this, aggression measures need to demonstrate and access *clinical cut-offs*…that is, the degree to which stimuli propel individuals outside of "normal" aggression levels into abnormal or harmful aggression. Some well-validated clinical measures such as the *Child Behavior Checklist* have such clinical cut-offs, whereas far too many other aggression measures including laboratory measures like

the TCRTT or surveys such as the *Aggression Questionnaire* have no clinical cut-offs. Interpretation of results from measures lacking clinical cut-offs is difficult, as it is not possible to evaluate whether individuals are demonstrating "harmful" levels of aggressive behavior.

Assessment measures for aggression used in video game violence studies fall into several categories, which are briefly discussed below. Recommendations for their use (or non-use) are provided. The order below roughly corresponds to the relative validity and degree to which they can be generalized to real-world aggression.

Laboratory based aggression paradigms. These measures, such as the TCRTT have been evolved to attempt to demonstrate causal models of aggression in laboratory settings. The validity of such measures has historically been controversial (Ferguson & Rueda, 2009; Savage, 2004; Tedeschi & Quigley, 2000) and their use is advised only with great caution, particularly when generalizing results to serious real-world acts of aggression or violence.

Teacher/Peer/Parent ratings. These survey instruments attempt to provide "outside" ratings of aggressiveness in order to avoid social desirability or other biasing effects. These range from highly valid report instruments such as the *Child Behavior Check-list* and similar clinical instruments, to poorly constructed ad-hoc measures that may measure "naughtiness" rather than actual aggression (see Savage, 2004 for a discussion). Although they may avoid self-report bias, other forms of bias may intrude instead, with such ratings devolving into "popularity" contests, particularly in relation to peer nominations of aggression. The validity of ad-hoc teacher, parent and peer ratings have generally been quite poor (Henry & Metropolitan Area Child Study Research Group, 2006), although the quality of validated clinical measures may be much higher.

Cognitive tests. These tests are largely ad-hoc laboratory measures which attempt to evaluate the "accessibility" of "aggressive thoughts." Such tests include word blanks, Implicit Association tests, reaction time tests, etc. Validity for these measures as predictors of real-world violence is almost universally poor.

Hypothetical aggression tests. These survey instruments such as the *Aggression Questionnaire* commonly ask participants how they might respond to a particular provoking stimuli. For instance survey items may ask "When people criticize me I feel like yelling at them." The survey doesn't ask about actual behaviors, but rather how the participant might respond to hypothetical situations. However, individuals' beliefs about how they *might* respond may not predict how they *actually* respond. As a class, such measures may be a better measure of aggressive personality rather than aggressive behavior. Evidence suggests that these measures are not great predictors of actual aggressive acts (Ferguson & Rueda, 2009).

Arrest records. These are actually seldom used, owing to the practical and legal barriers to securing them. These are probably reasonably valid outcome measures

when available, although it must be remembered that not all violent acts results in arrest.

Self-reported violent crimes.   Several well-used and validated measures such as the *Negative Life Events* questionnaire and *National Youth Survey* have been used to evaluate the incident of violent criminal acts. These are generally good instruments, although social desirability naturally is a significant concern.

Clinical measures.   These include well-researched and validated instruments such as the *Child Behavior Checklist, Personality Inventory for Youth* and *Olweus Bullying Questionnaire*. These instruments tend to have a plethora of validity data and, as such, are amongst those most highly recommended for use as outcome measures.

Actuarial measures.   With little question, the best predictors of violence risk are actuarial measures developed specifically for this purpose (i.e. *HCR-20, VRAGS, Defendant and Offender Screening Test)*. Such measures have good validity coefficients regarding violence prediction (although violence prediction always remains difficult), and clear clinical cut-offs. Unfortunately the expertise required in administrating and scoring these tests often precludes their use in video game research, and to date no video game studies have used these excellent measures.

Measuring Violence in Video Games

Most of the controversy over measurement in video game studies is regarding the measurement of aggression. Yet, the measurement of violence in video games itself is not without concern. Three basic systems for rating the amount of violence in games have been used in research. Each has their own set of strengths and potential problems.

   First, researchers may attempt to engage in a content analysis of popular video games, using standardized criteria to rate them for violent content. So long as content ratings carefully adhere to an objective rubric for scoring the games, this approach has some intuitive appeals. Exposure to violence in games would be consistently measured across individuals. However several issues limit the widespread use of this method. First, although this approach does succeed in creating a standardized criteria for measuring video game violence across individuals, there's no means of assuring that the ratings themselves are valid. The initial rubric itself, though standardized, may simply be a product of the researcher's personal opinions and biases and not represent video game violence in any "true" sense. This may particularly be an issue with older researchers who are not particularly familiar with video game play (that games such as *Zaxxon* have been labeled as highly violent, highlights this issue). Second, and perhaps more critical, in the early era of video games, when the number of titles produced was relatively few, this approach may have made sense. However, with the rapid turnover of modern games, developing a meaningful metric for existing games may simply be impractical, as any existing database is quickly rendered obsolete.

A second option, and probably that most commonly used, is to ask for participants to estimate the amount of violence themselves in the games they play. As this is the easiest approach, its common use is understandable. However, this does limit consistency in ratings between individuals as differing participants may have differing perspectives about what constitutes violence in video games.

The third approach is to rely on ratings and content descriptors provided by the Entertainment Software Ratings Board (ESRB). The ESRB rates the age appropriateness of all video games commercially sold. The ESRB ratings increasingly have been supported by the FCC, Parents Teachers Association, numerous governors and attorneys general, and even anti-game activist groups such as the National Institute of Media and the Family. As such the ESRB ratings are generally viewed as highly valid, and are available for all commercial games sold. However, using this system does not allow for measurement of violence in non-commercial games such as those sold or provided free over the internet. Also, games may receive higher ratings (that is, more adult ratings) for reasons other than violent content (such as for language, sexual content, etc.) Thus, ESRB ratings are not a perfect estimate of violent content.

Remarkably little research has examined which of these three approaches are best for estimating violence in video games. It would be advisable for future research to examine the comparative utility of these methods for measuring video game violence.

Implications for Serious Games

Several implications for serious games bear discussing in relation to assessments of video game violence and aggression. These regard both the inclusion of violence in serious games, as well as lessons learned from the aggression literature itself.

It is well understood that violent video games are popular among even young children (Kutner & Olson, 2008). Research on video game violence suggests that, contrary to popular beliefs, harmful effects of video game violence exposure are small to nil. Given that serious games (games intended for an ultimate cause other than to simply entertain) are faced with the prospect of competing with standards of quality and game play found in commercial games, including violent content in serious games may have certain appeals. In particular, including violent content (at least mild violence…there are no metals for getting banned in Australia) may help attract young players to the game so that the "serious" content of the game can be more faithfully delivered.

Of greater issue, perhaps is that of "lessons learned." To be frank, the video game violence literature has mangled assessment to such an extent that repairing the damage done would take years of effort even if a more sophisticated program of assessment and interpretation were undertaken by the field at large which, sadly, appears unlikely for the near future. However the same issues seen in video game violence assessments, could easily become problematic for assessments related to serious games. Just as the video game violence literature has become politicized, so too may those who spend months or years developing serious games become heavily and emotionally invested in demonstrating the positive value of those games. Evaluators of serious games, invested emotionally, temporally and financially in the games, may have difficulty employing objective assessments in the evaluation of

said games. As a consequence, the use of ad hoc, poorly validated measures may become just as prevalent as with video game violence research.

Several thoughts are offered in hopes that they may help guide serious games researchers in assessing outcomes associated with games:

The use of standardized and valid measures is not optional. First and foremost serious games evaluators can avoid the pitfalls of video game violence research by insisting on the use of standardized and well-validated assessment instruments early on. While investigators themselves can take care to use well-developed outcome measures, peer-reviewers also have great responsibility in assuring that assessments are valid and objective.

Best outcome measures will be developed by groups OTHER than the group evaluating the serious game. There is a risk that evaluators of serious games may tailor outcomes too closely to the content of the game rather than to the desired outcome. For example, consider a developer of a video game meant to teach biology to middle school students. In developing an outcome measure, researchers invested in the game may be tempted (consciously or unconsciously) to develop outcomes that tap into the content covered in the game, rather than human biology knowledge content more generally. Outcomes developed in such a manner would obviously be biased and grossly overestimate the effectiveness of serious games. Outcomes that are developed more distally and tied to objective outcomes of value to society rather than game developers are of the highest quality.

Conclusions are limited by the quality of assessments. One of the biggest concerns about the violent video game literature has been the willingness of too many researchers to ignore the limits of aggression assessments when generalizing results to serious acts of real-world aggression or violence. Evaluators of serious games must take great care not to repeat this mistake. Although grand pronouncements make for excellent press releases, too often time reveals their limitations. Careful, more humble pronouncements of success, and honest appraisals of failures will bear more fruit long-term.

In conclusion, there is much that the emerging field of serious games can learn from the past mistakes of the video game research field. With caution, care and close attention to well-established principles of measurement, the field of serious games can shine as an example for proper use of assessments in outcome research.

REFERENCES

Anderson, C. (2004). An update on the effects of playing violent video games. *Journal of Adolescence*, *27*, 113–122.

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-analysis. *Psychological Science*, *12*, 353–359.

Anderson, C., & 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.

Anderson, C., & Ford, C. (1987). Affect of the game player: Short term effects of highly and mildly aggressive video games. *Personality and Social Psychology Bulletin*, *12*(4), 390–402.

Anderson, C., & Murphy, C. (2003). Violent video games and aggressive behavior in young women. *Aggressive Behavior*, *29*, 423–429.

Barlett, C. P., & Anderson, C. A. (2009). Violent video games and public policy. Chap. in T. Bevc & H. Zapf (Eds.), *Wie wir spielen, was wir werden: Computerspiele in unserer Gesellschaft* (pp. 227–240). Konstanz: UVK Verlagsgesellschaft.

Baron, R., & Richardson, D. (1994). *Human aggression*. New York: Plenum Press.

Bartholow, B., Bushman, B., & Sestir, M. (2006). Chronic violent video game exposure and desensitization to violence: Behavioral and event-related brain potential data. *Journal of Experimental Social Psychology*, *42*, 532–539.

Carnagey, N., & Anderson, C. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition and behavior. *Psychological Science*, *16*, 882–889.

Colwell, J., & Kato, M. (2003). Investigation of the relationship between social isolation, self-esteem, aggression and computer game play in Japanese adolescents. *Asian Journal of Social Psychology*, *6*, 149–158.

Cooper, J., & Mackie, D. (1986). Video games and aggression in children. *Journal of Applied Social Psychology*, *16*(8), 726–744.

Dominick, J. (1984). Videogames, television violence and aggression in teenagers. *Journal of Communication*, *34*(2), 136–147.

Durkin, K., & Barber, B. (2002). Not so doomed: Computer game play and positive adolescent development. *Applied Developmental Psychology*, *23*, 373–392.

ESA, VSDA and IRMA v. Blagojevich, Madigan and Devine. (2005). Case No. 05 C 4265.

Ferguson, C. J. (2009). Media violence effects: Confirmed truth, or just another X-File? *Journal of Forensic Psychology Practice*, *9*(2), 103–126.

Ferguson, C. J. (2008). The school shooting/violent video game link: Causal link or moral panic? *Journal of Investigative Psychology and Offender Profiling*, *5*, 25–37.

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior*, *12*, 470–482.

Ferguson, C. J., Averill, P. M., Rhoades, H., Rocha, D., Gruber, N., & Gummattira, P. (2005). Social isolation, impulsivity and depression as predictors of aggression in a psychiatric inpatient population. *Psychiatric Quarterly*, *76*(2), 123–137.

Ferguson, C. J., & Beaver, K. M. (2009). Natural born killers: The genetic origins of extreme violence. *Aggression and Violent Behavior*, *14*(5), 286–294.

Ferguson. C. J., & Kilburn, J. (2009). The Public health risks of media violence: A meta-analytic review. *Journal of Pediatrics*, *154*(5), 759–763.

Ferguson, C. J., & Rueda, S. M. (in press). The Hitman study: Violent video game exposure effects on aggressive behavior, hostile feelings and depression. *European Psychologist*.

Ferguson, C. J., & Rueda, S. M. (2009). Examining the validity of the modified Taylor competitive reaction time test of aggression. *Journal of Experimental Criminology*, *5*(2), 121–137.

Ferguson, C. J., Rueda. S., Cruz, A., Ferguson, D., Fritz, S., & Smith, S. (2008). Violent video games and aggression: Causal relationship or byproduct of family violence and intrinsic violence motivation? *Criminal Justice and Behavior*, *35*, 311–332.

Ferguson, C. J., San Miguel, C., & Hartley, R. D. (2009). A multivariate analysis of youth violence and aggression: The influence of family, peers, depression and media violence. *Journal of Pediatrics*, *155*(6), 904–908.

Freedman, J. (1992). Television violence and aggression: What psychologists should tell the public. In P. Seudfeld & P. Tetlock (Eds.), *Psychology and social policy*. New York: Hemisphere Publishing Corp.

Funk, J., Hagan, J., Schimming, J., Buchman, D., Myers, M., & Bullock, W. (2002). Aggression and psychopathology in adolescents with a preference for violent electronic games. *Aggressive Behavior*, *28*(2), 134–144.

Gentile, D., Lynch, P., Linder, J., & Walsh, D. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors and school performance. *Journal of Adolescence*, *27*, 5–22.

Grimes, T., Anderson, J., & Bergen, L. (2008). *Media violence and aggression: Science and ideology*. Thousand Oaks, CA: Sage.

Grossman, D. (1996). *On killing: The psychological cost of learning to kill in war and society*. Boston: Back Bay Books.

Hawley, P., & Vaughn, B. (2003). Aggression and adaptive function: The bright side to bad behavior. *Merrill-Palmer Quarterly*, *49*, 239–242.

Henry, D., & Metropolitan Area Child Study Research Group. (2006). Associations between peer nominations, teacher ratings, self-reports, and observations of malicious and disruptive behavior. *Assessment*, *13*, 241–252.

Huesmann, L. R. (2007). The impact of electronic media violence: Scientific theory and research. *Journal of Adolescent Health*, *41*, S6–S13.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS.Med*, *2*, e124. Retrieved July 14, 2009, from http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

Kirsh, S. (1998). Seeing the world through Mortal Kombat-colored glasses: Violent video games and the development of a short-term hostile attribution bias. *Childhood: A Global Journal of Child Research*, *5*(2), 177–184.

Konijn, E. A., Nije Bijvank, M., & Bushman, B. J. (2007). I wish I were a warrior: The role of wishful identification in effects of violent video games on aggression in adolescent boys. *Developmental Psychology*, *43*, 1038–1044.

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

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

Olson, C. K. (2010). Children's motivations for video game play in the context of normal development. *Review of General Psychology*.

Olson, C. (2004). Media violence research and youth violence data: Why do they conflict? *Academic Psychiatry*, *28*, 144–150.

Queller, & Strassman (2002). Quick guide: Kin selection. *Current Biology*, *12*, R832.

Ritter, D., & Eslea, M. (2005). Hot sauce, toy guns and graffiti: A critical account of current laboratory aggression paradigms. *Aggressive Behavior*, *31*, 407–419.

Savage, J. (2004.) Does viewing violent media really cause criminal violence? A methodological review. *Aggression and Violent Behavior*, *10*, 99–128.

Sherry J. (2007). Violent video games and aggression: Why can't we find links? In R. Preiss, B. Gayle, N. Burrell, M. Allen, & J. Bryant, (Eds.), *Mass media effects research: Advances through meta-analysis (*pp. 231–248). Mahwah, NJ: L. Erlbaum.

Silvern, S. B., & Williamson, P. A. (1987). The effects of video game play on young children's aggression, fantasy and prosocial behavior. *Journal of Applied Developmental Psychology*, *8*, 453–462.

Smith, J. M. (1964). Group selection and kin selection. *Nature*, *201*, 1145–1147.

Smith, P. (2007). Why has aggression been thought of as maladaptive? In P. Hawley, T. Little, & P. Rodkin (Eds.), *Aggression and adaptation: The bright side to bad behavior* (pp. 65–83). Mahwah, NJ: Lawrence Erlbaum.

Tedeschi, J., & Quigley, B. (1996). Limitations of laboratory paradigms for studying aggression. *Aggression & Violent Behavior*, *2*, 163–177.

Tedeschi, J., & Quigley, B. (2000). A further comment on the construct validity of laboratory aggression paradigms: A response to Giancola and Chermack. *Aggression & Violent Behavior*, *5*, 127–136.

Thompson, J. (2007). *Massacre at Virginia Tech: Interview with MSNBC*. Retrieved July 16, 2008, from http://www.msnbc.msn.com/id/18220228/

Thompson, K., & Haninger, K. (2001). Violence in E-rated video games. *Journal of the American Medical Association*, *286*(5), 591–598.

Unsworth, G., Devilly, G., & Ward, T. (2007). The effect of playing violent videogames on adolescents: Should parents be quaking in their boots? *Psychology, Crime and Law*, *13*, 383–394.

van Schie, E., & Wiegman, O. (1997). Children and videogames: Leisure activities, aggression, social integration, and school performance. *Journal of Applied Social Psychology*, *27*(13), 1175–1194.

Williams, D., & Skoric, M. (2005). Internet fantasy violence: A test of aggression in an online game. *Communication Monographs*, *72*, 217–233.

World Health Organization. (2002). *World report on violence and health*. Geneva, Switzerland: World Health Organization.

Christopher J. Ferguson
Texas A&M International University