Measuring Morality in Videogames Research

Malcolm Ryan

Department of Computing
Macquarie University
Sydney Australia
malcolm.ryan@mq.edu.au

Paul Formosa

Department of Philosophy
Macquarie University
Sydney Australia
paul.formosa@mq.edu.au

Stephanie Howarth

Department of Cognitive Science
Macquarie University
Sydney Australia
stephanie.howarth@mq.edu.au

Dan Staines

Department of Communication Studies
Concordia University
Montreal Canada
dan@danstaines.com

ABSTRACT

There has been a recent surge of research interest in videogames of moral engagement for entertainment, advocacy and education. We have seen a wealth of analysis and several theoretical models proposed, but experimental evaluation has been scarce. One of the difficulties lies in the measurement of moral engagement. How do we meaningfully measure whether players are engaging with and affected by the moral choices in the games they play? In this paper, we survey the various standard psychometric instruments from the moral psychology literature and discuss how they might be applied in the evaluation of games.

Keywords

moral psychology; ethics; psychometric testing; videogame design.

1. INTRODUCTION

Morally significant videogames (for entertainment, education and advocacy) are becoming commonplace (Zagal 2011). The academic response to this rising tide has produced a wealth of analysis and a variety of theoretical models for the design of such videogames (Belman and Flanagan 2010; Flanagan et al. 2007; Ryan et al. 2017; Schrier 2015; Sicart 2013). However, so far there has been little experimental validation of these models. One reason for this is the difficulty in rigorously measuring the moral qualities we are interested in. How do a player's moral values affect the way they play a game? And can playing a game influence a player's real-world moral behaviours? To find better-than-anecdotal answers to these questions, we need reliable measures for identifying correlations between morality and gameplay. To this end, we turn to the field of moral psychology. Moral psychologists have asked questions like this for close to a century and have developed a collection of psychometric instruments to answer them. The value of these instruments for game researchers is that they are standardised and reproducible, grounded in theory, and validated through empirical testing. While these instruments are not uncontroversial, they have much to recommend over the ad-hoc methods games researchers might otherwise use.

In this paper, we review several of the myriad available tests described in the moral psychology literature. Rather than attempt to include every such test,[1] we instead focus on those most relevant and useful for games researchers. To show how to employ these tests effectively, we first describe some of the important background considerations in selecting an instrument for games research, before introducing the instruments themselves. We then highlight tests that have already been used in videogames research and suggest how other tests might be employed in future research.

2. MORALITY AND GAMES RESEARCH

There are two basic questions we might ask about games and morality:

1. How does a player's morality affect the way they play games?
2. How do the games they play affect a player's morality?

The first of these questions regards the player's *moral engagement*, i.e. how the player utilises their real-world moral capacities in gameplay. The second question regards *moral effects*, i.e. the effects gameplay has on real-world morality. These effects can be very short term, lasting only a few minutes, or long term, in which case they affect a player's *moral development* and underlying moral functions. The game researcher's first task is to identify which of these questions they wish to investigate and, if the latter, whether they are interested in short- or long-term effects. Psychologists largely agree that changes in moral functioning occur over long periods of time (years or decades), if at all. It is extremely unlikely that a single session of gameplay would have any lasting effect on a player's moral development, even if it has a short-term effect on a particular trait, such as aggression (Anderson et al. 2010; Ferguson 2007). Of the tests we present here only a few, such as the *Defining Issues Test*

[1] For a more comprehensive catalogue of over 300 measures used in behavioural ethics, see Agle et al. (2014).

(DIT) and *Measure of Prosocial Reasoning* (PROM), have enough longitudinal data to demonstrate development and how designed interventions and life events (such as completing postgraduate education) affect its trajectory. Answering research questions regarding moral development and gameplay requires similar longitudinal studies of players over many years.

Relevant instruments must then be incorporated into a well-defined research question and experimental design. There are many experimental design variations that can be employed to answer different empirical problems. Typical structures involve the identification of independent and dependent variables, and the use of experimental and control groups. The experimental group typically undergoes some form of intervention that the control group does not, such as playing a videogame, and then the two (or more) groups are compared to examine the relationship between the independent and dependent variables. Randomly assigning people to relevant groups and ensuring sufficient sample sizes all assist in ensuring adequate internal validity, which is essential for determining cause-effect relationships. In games research, the experimental structure can often be broken down into one or more of the following three phases:

- **Pre-test** measures made before playing the game, to establish the psychological characteristics of the player.
- **Response** measures made while playing the game, to measure in-game choices and behaviours.
- **Post-test** measures made after play (immediately or after a time delay), to measure whether the game affected the player in the short- or long-term.

2.1 Theoretical Underpinnings

After formulating a research question, the game researcher then needs to be aware that each instrument is embedded in a particular theory and its usage brings with it associated theoretical commitments. There are a variety of psychological theories to explain moral functioning, from Kohlberg's cognitive developmental approach (Kohlberg 1981), to social cognitivism (Lapsley and Narvaez 2005) and social intuitionism (Haidt and Bjorklund 2007), although it is beyond the scope of this paper to explore these in detail. The Moral Judgment Interview (MJI), for example, is derived from a model of moral judgment that privileges explicit verbalisation and conscious deliberation. Alternatively, the Moral Foundations Questionnaire (MFQ) is based on a social intuitionist model of morality that privileges spontaneous "gut feelings" and largely relegates conscious deliberation to the role of post-hoc rationalisation.

To understand why these different facets are privileged by different instruments, we follow Lapsley and Hill (2008) in subscribing to a dual-process model of moral cognition which acknowledges two types of reasoning processes: Type 1, which is fast, implicit and intuitive; and Type 2, which is slow, explicit and deliberative (Evans and Stanovich 2013). As mentioned above, some theories prioritise deliberate reasoning (Type 2) while others emphasise intuition and

emotion (Type 1) when examining moral judgment. Therefore, measurement should correspond with the theory it supports while being aware of the limitations or biases it may encourage. For example, our conscious minds do not have direct access to the operation of our intuitive processes and so asking a subject to evaluate a decision made intuitively through explicit self-report measures, such as interviews and surveys, could result in post-hoc rationalisation (Pennycook et.al. 2015), especially if the person is a novice at making the kinds of judgments required by the situation. Rationalisation, or analytical thinking, has been related to less traditional moral values (Pennycook et al. 2015) and can also be prone to the effects of social desirability (Nederhof 1985)."

Moral behaviour involves several different skills. Game researchers need to know which they are interested in investigating in order to select the appropriate instrument. The Four Component Model (Rest 1983) identifies four broad categories of cognitive-affective processes involved in a moral act:

1. **Moral Focus** – the extent to which one is committed to one's moral choices and prioritises moral concerns over other concerns.
2. **Moral Sensitivity** – the ability to identify morality in the real world, to understand the motivations of others, and to perceive the consequences of one's behaviour.
3. **Moral Judgment** – the ability to understand moral concepts, reason about moral issues, and make moral judgments.
4. **Moral Action** – the ability to overcome temptations, persist in the face of adversity, be effective in taking action, and do the right thing even when it's hard.

Different psychometric tests focus on different components. Some assess the degree to which subjects prioritise moral concerns (moral focus), some ask the subject to recognise the moral importance of a situation (moral sensitivity), some ask for a reasoned moral argument in response to a situation (moral judgment), and others simulate a morally significant situation and look for moral behaviour (moral action). These distinctions are unavoidably interrelated– real moral decisions involve all four areas, often in parallel – but there are still important differences. For example, one might be capable of understanding complex moral arguments (moral judgment) but lack the courage required to translate this into behaviour (moral action).

A further theoretical consideration is whether an instrument includes "macro-" or "micro-moral" decisions (Narvaez and Bock 2002). Macro-morality deals with large scale societal issues, such as the legal permissibility of abortion, while micro-morality deals with personal problems, such as whether to lie to a friend. Popular tests (e.g. the DIT and MFQ) favour macro-moral issues as they are reasonably universal. Micro-moral issues tend to be more domain-specific and so require the design of special-purpose instruments (Thoma 2014). This is relevant for game researchers insofar as design features associated with specific genres impact how games represent the scope of moral choice. The sim genre, including games like the *McDonalds Videogame* (Molleindustria 2006) and the *Civilization* series (Microprose 1991), is typically better suited to representing macro-moral choice from a 'god's eye view', whereas RPG and adventure

games, such as *The Walking Dead* (Telltale Games 2012), may be more appropriate for representing interpersonal, micro-moral scenarios, for a specific player-character.

2.2 Instrument Validity and Reliability

For any scientific instrument to provide measures that can be trusted, it must be established as both valid and reliable. Reliability is the degree to which an assessment tool is consistent in the results it produces, including both test-retest reliability and internal consistency between different items on a test (Samuels 2015). Validity refers to how well a test measures what it is supposed to measure. *Internal* validity refers to how well the experiment is conducted, in that no other confounding variables could explain the findings. *External* validity represents how well the findings can be generalised outside the laboratory (*ecological* and *population* validity). In terms of scale development, researchers must confirm *construct validity*, i.e. whether the instrument or scale measures the theoretical concept it claims to measure. This requires confirmation of both *convergent* (items in a scale are related when they should be) and *divergent* validity (items in the scale are not related when they shouldn't be). To establish *content validity*, a scale measuring moral focus, for example, must adequately measure all aspects of what constitutes that construct. *Criterion-related validity* relates to other measures that are correlated in some way to the construct, and can be separated into *predictive* and *concurrent* validity, i.e. does the scale predict what it theoretically should be able to predict, and does it differentiate between the groups it should be able to distinguish between. Since we cannot discuss the details of the individual validity and reliability of each measure outlined here, we direct the reader to the individual papers referenced below for those details, although we sometimes mention when an instrument is particularly well validated.

In addition to reliability and validity, in deciding to use any of these instruments, game researchers must also consider pragmatic issues, particularly the time and effort required to administer and score the test. Some tests, such as the MJI, require in-depth training for experimenters. Others can be taken online and scored automatically. Where relevant, we give an indication of the effort involved in using an instrument.

2.3 Instrument Types

While pre- and post-test measures might involve behavioural observation (such as seeing whether a subject helps a confederate pick up her dropped books) most game research will need to employ different types of self-report measures in addition to in-game response measures. These can take two main forms (Thoma 2014):

- **Production tasks**, in which the subject is asked open-ended questions to verbally explain their reasoning for a given problem, such as through a verbal interview or a written survey answer.

- **Evaluation tasks**, in which the subject is asked to evaluate their reaction to a set of items, typically using a Likert-scale to assign each item a rating (e.g. 1 = Strongly Agree" to 5 = "Strongly Disagree")[2]

Production tasks test a subject's ability to spontaneously generate moral insights, while evaluation tasks test their ability to recognise and evaluate morally significant items, such as statements and vignettes (Thoma 2014).

The structure of the instrument affects what it measures, particularly with regard to Type 1 and Type 2 processes. Production tasks invite conscious deliberation to verbally construct explanations. As such, they primarily engage Type 2 processes. Evaluation tasks only require the subject to evaluate a stimulus, not verbally explain their evaluation, and so they can engage either intuitive Type 1 or effortful Type 2 responses (Gibbs et al. 1982; Thoma 2014). Such measures do not determine whether a subject will or will not override or rationalise their initial intuitive response before answering.

Both production and evaluation tasks have difficulty distinguishing intuitive (Type 1) moral processes from deliberative (Type 2) processes. Everyday morality employs a mixture of both, but if we want to examine moral intuition in isolation, verbal interviews and pencil-to-paper tasks can be problematic. Adding cognitive load, by increasing time pressure or including a secondary task, can help expose Type 1 responses by engaging Type 2 cognitive resources (Greene et al. 2008).

Another approach is to make use of *priming* (Bargh and Chartrand 2000). The technique usually focuses on exposing a subject to one type of stimulus to influence their response to another stimulus without conscious awareness. The Affect Misattribution Procedure (AMP) is a technique that exploits this priming effect (Payne and Lundberg 2014). In its general form, the experimenter creates two sets of stimuli, typically words or images. The first set contains the stimuli for which we want to measure a response, and the second set of stimuli are ambiguous (typically simple abstract images, such as Chinese pictographs for an audience who do not know their meaning). Experimental evidence shows that the subject's intuitive affective response to the first stimulus is strongly reflected in their evaluation of the second, ambiguous stimulus. Priming can also influence moral judgment. For example, Cameron et al. (2013) showed that a disgusting image can influence the strength of moral judgments of counter-normative cultural practices for unskilled emotional differentiators even when subjects were warned about this possible influence.

Another issue with self-report measures is subjects' concern for the social desirability of their answers, which can result in self-censorship (Nederhof 1985). In socially-sensitive areas, such as morality, subjects tend to respond in ways they believe will be more socially acceptable, biasing results. Further, self-reports about what a person believes they would do in hypothetical situations may not correspond with what they would do in real-world scenarios (Krebs et

[2] While we use the term "Evaluation Task" to describe instruments like the DIT, they are often referred to in the literature (e.g. Thoma, 2014) as measuring "recognition data". We use the former because it is the least ambiguous of the two and, we believe, most accurately captures the nature of the described instruments.

al. 1997). This points to an important gap between hypothetical moral judgment and real-world moral action (Blasi 1980).

Beyond employing various instruments in pre- and post-tests, game researchers can also measure what players *do* while playing a videogame. These are:

- **Response measures**, in which the in-game responses and behaviours of a subject are measured.

These in-game behaviours and choices can be affected by interventions, such as adding time pressure, or through modification of the gaming experience by changing game mechanics, player choice options, or adding the facility to customize avatars. Game researchers can also measure response dynamics, such as reaction times or mouse-movements, to provide insight into the underlying cognitive processes involved, as discussed below.

3. MORALITY MEASURES

There are a wide variety of morality measures game researchers might use (Agle et al. 2014; Jordan 2007). We describe twenty such tests here, as summarised in Table 1. Rather than attempt an exhaustive (and overly lengthy) list, we have instead focused on tests that satisfy these criteria:

- **Theoretical foundation**: tests that are backed by well-developed theory
- **Validity**: tests that have been evaluated for validity
- **General purpose**: tests that are not specialised to a particular domain
- **Availability**: tests that are readily available to researchers
- **Wide use**: tests that have been widely used by researchers
- **Applicability**: tests that are relevant to videogames researchers

We present these tests grouped, initially, by task type (first Production Tasks and then Evaluation Tasks), and secondly, we have grouped Evaluation Tasks by theoretical foundation before, thirdly, examining response measures.

We have also labelled each test according to its focused moral component, although due to the overlap between components this can be contentious. We generally classify instruments that measure what is most important to a person as a test of moral focus, and instruments that require subjects to make and justify particular judgments as tests of moral judgment, although there is clearly overlap between the two. A glance at the table will reveal a preponderance of moral judgment and (more recently) moral focus measures, with only a few moral sensitivity measures, reflecting the historical focus of research. There is a significant lack of general-purpose measures of moral action in part due to the difficulty of measuring it. This is an avenue where games research could have a significant impact by providing a virtual context in which various moral choices can be put into action, with varying degrees of resistance, without negative real-world implications.

Test Name	**Format**	**Theoretical foundation**	**Focused moral component**	**References**

Moral Judgment Interview (MJI)	Production task Interview	Cognitive developmentalism	Moral judgment	(Colby and Kohlberg 1987) (Narvaez and Bock 2002)
Socialmoral Reflection Measure (SRM)	Production task Survey	Cognitive developmentalism	Moral judgment	(Gibbs et al. 1982)
Socialmoral Reflection Measure – Short Form (SRM-SF)	Production task Survey	Cognitive developmentalism	Moral judgment	(Basinger, Gibbs, and Fuller 1995)
Moral Competence Test (MCT)	Evaluation task Survey	Cognitive developmentalism	Moral judgment	(Lind 2013)
Defining Issues Test (DIT)	Evaluation task Survey	Social cognitivism	Moral judgment	(Rest 1979)
Defining Issues Test 2 (DIT-2)	Evaluation task Survey	Social cognitivism	Moral judgment	(Rest, et al. 1999B)
Moral Foundations Questionnaire (MFQ)	Evaluation task Survey	Social intuitionism	Moral focus	(Graham et al. 2011)
Moral Foundations Vignettes (MFV)	Evaluation task Survey	Social intuitionism	Moral sensitivity	(Clifford et al. 2015)
Moral Foundations Sacredness Survey (MFSS)	Evaluation task Survey	Social intuitionism	Moral focus	(Graham and Haidt 2012)
Measure of Prosocial Moral Reasoning (PROM)	Evaluation task Survey	Prosocial Morality	Moral judgment	(Carlo, Eisenberg, and Knight 1992)
Measure of Moral Orientation (MMO)	Evaluation task Survey	Moral identity	Moral focus	(Liddell and Davis 1992)
Measure of Moral Identity (MMI)	Evaluation task Survey	Moral identity	Moral focus	(Aquino and Reed 2002)

Triune Ethics Orientation (TEO)	Evaluation task Survey	Triune ethics metatheory	Moral focus	(Narvaez and Hardy 2016)
Schwartz Value Scale (SVS)	Evaluation task Survey	Value theory	Moral focus	(Schwartz 2012)
Portrait Values Questionnaire (PVQ)	Evaluation task Survey	Value theory	Moral focus	(Schwartz 2012)
Ethics Position Questionnaire (EPQ)	Evaluation task Survey	Universalism vs. relativism & idealism vs. realism	Moral judgment	(Forsyth 1980)
Measure of Ethical Viewpoints	Evaluation task Survey	Formalism vs Consequentialism	Moral judgment	(Brady and Wheeler 1996)
Moralization of Everyday Life (MELS)	Evaluation task Survey	Micro-morality	Moral judgment	(Lovett, Jordan, and Wiltermuth 2012)
Multiple Intelligence Profiling Questionnaire Ethical Sensitivity Scale (MIPQ-ESS)	Evaluation task Survey	Self-assessed moral sensitivity	Moral sensitivity	(Tirri and Nokelainen 2011)
Immediate Affect Toward Moral Stimuli (IAMS)	Affect Misattribution Procedure	Dual Process theory	Moral sensitivity	(Hoffman and Baumert 2010)

3.1 Production Tasks

For much of the 20th Century, moral psychology was dominated by cognitive developmentalism: a rationalist, justice-oriented conception of moral functioning championed by Kohlberg. The centrepiece of Kohlberg's work was the MJI. Reflecting cognitive developmentalism's theoretical commitments, the MJI is a production task that emphasises explicit (Type 2) verbal moral reasoning.

Moral Judgment Interview (MJI)

Following Piaget, Kohlberg was interested in discovering structural patterns in how subjects arrive at and justify moral judgments. He recognised six distinct patterns which he claimed form a "universal invariant" sequence of stages of moral development, from stage 1 "heteronomous morality" through to stage 6 "universal ethical principles" (Colby and Kohlberg 1987). The MJI attempts to

establish a subject's stage of development through a structured interview. The interviewer presents three hypothetical ethical dilemmas and asks the subject to make a judgment about what action should be taken. Each dilemma is followed by a series of probing questions to elicit the subject's ethical justifications for their judgment. A scoring manual (Colby et al. 1987) is provided, containing an extensive list of possible responses and relating them to each of the six stages.

Sociomoral Reflection Measure (SRM and SRM-SF)

The SRM was designed to be a shorter alternative to the MJI (Gibbs et al. 1982). Like the MJI, it assesses reflective moral judgments about what action should be taken based on Kohlberg's six stages. However, it simplifies the interview process to a pen-and-paper questionnaire in which the subject writes open-ended answers justifying their decisions. The ethical dilemmas of the MJI and SRM have been criticised for being culturally specific (Gibbs et al. 2007). This criticism led to the development of the SRM-SF (Basinger et al. 1995) which replaces the moral dilemmas with simpler lead-in statements. Each statement is followed by an evaluation question and an open-ended production task where the subject justifies their answer. These responses must be coded and scored to yield an overall Moral Maturity Score in the range 1 to 4, corresponding to the first four stages of Kohlberg's sequence.

Guidelines for use: These instruments measure moral judgment by looking at the level of reasoning at which the subject operates. The MJI is mainly of historical interest, as it is extremely labour-intensive, and researchers require significant training to conduct and code interviews reliably (Gibbs et al. 1982). While the SRM is simpler to implement, significant training is still required to score it reliably and coding remains difficult and time-consuming; furthermore, the shorter SRM-SF cannot measure the highest two stages described by Kohlberg's theory. The reliance of all these tests on explicit verbally produced responses is also considered a weakness (Rest et al. 1997A), given the importance of intuitive (Type 1) moral reasoning and the difficulty of verbalising tacit knowledge (Polanyi 1966). Moreover, the invariant sequence of stages of moral development has been called into doubt (Lapsley and Narvaez 2005) and has been replaced by the schema theory of Rest and colleagues outlined below.

3.2 Evaluation Tasks

Despite enjoying a great deal of empirical support, cognitive developmentalism suffers from a variety of philosophical and empirical shortcomings (Rest et al. 1999C) and is no longer regarded as moral psychology's dominant paradigm (Vozzola 2014). In its wake, several promising alternatives have arisen, each with their own instruments for measuring moral functioning and development. These measures tend to involve evaluation tasks, which are easier to administer, better able to assess Type 1 responses, and less dependent on a subject's verbal skills than production tasks.

3.2.1 Neo-Kohlbergian

Defining Issues Test (DIT and DIT-2)

Developed by Rest (1979) and colleagues, DIT (or DIT-1) is a survey of moral judgment that retains the core insights of cognitive developmentalism while jettisoning its more problematic elements. Instead of Kohlberg's strict six stage sequence, Neo-Kohlbergians (including the social cognitivists) recognise three schemas (patterns of reasoning) used by individuals to process moral scenarios and produce moral judgments (Rest et al. 1999C). These are the "Personal Interest" (e.g. self-interest), "Maintaining Norms" (e.g. uphold the law and obey social conventions) and "Post-Conventional" schemas (e.g. universal ethical ideals). An important point of difference between schemas and stages is that schemas are not developed in an invariant sequence of gradually increasing complexity. Instead, development is conceptualised as "shifting distributions, whereby the more primitive ways of thinking are gradually replaced by more advanced ways of thinking" (Rest et al. 1999A. p. 298). Where Kohlberg saw moral development as a metaphorical staircase, an apt metaphor for schema development is that of overlapping waves.

The DIT presents the subject with six macro-moral dilemmas and asks them to rate and rank lists of considerations that played a role in their decision making about what should be done. Considerations are presented as deliberately ambiguous to encourage subjects to 'fill in' missing details with their own moral schema (Narvaez and Bock 2002). The result of the test is a 'P-score' that indicates the subject's preference for the post-conventional schema, which is regarded as the highest level of mature moral judgment. A revised version of the test, the DIT-2, has been released (Rest, et al. 1999B). While the questions have been revised and shortened, the basic structure remains the same, with the main difference being the scoring scheme. A new index (the N2-index) is more reliable because it incorporates "the acquisition of new thinking (increases in P score)" as well as "systematic rejection of simplistic thinking (decreases in Stages 2 and 3)" (Rest et al. 1997B p.500). Recently, a related instrument has been developed for measuring, in adolescent populations, "intermediate concepts", which sit above the DIT's "bedrock concepts" defined by its schema scores (Thoma et al. 2013).

Moral Competence Test (MCT)

Initially developed by Lind (2013) in 1978, the MCT is designed to assess "moral competence" – described as the "ability to resolve problems or conflicts on the basis of reasoning" (Lind 2019). Although similar in some respects, moral competence differs from Kohlberg's moral judgment construct in that it is derived from Dual Aspect Theory, which (among other things) acknowledges the role of implicit reasoning in moral decision making. The MCT presents the subject with two moral dilemmas and a selection of arguments for and against possible actions, with one argument for each of Kohlberg's stages of moral development. The subject is asked to rate each argument in terms of its importance.

An important difference between the MCT and the DIT is in the scoring process. Lind claims that the DIT's P-score measures "moral preference" and is not a strong measure of moral maturity (Lind 2013). Instead, he proposes a C-score which measures the consistency with which the subject rates different staged

arguments, without preferring one stage over another. Part of the motivation for this is the claim that the DIT is biased to rate liberal morality higher than conservative morality, however the validity of these claims and the strength of the MCT overall have been strongly criticised (Rest et al. 1997A).

Guidelines for use: The DIT and DIT-2 measure a subject's preference for and capacity to account for post-conventional moral judgments, whereas the MCT C-score provides a measure of how consistently a subject operates at their moral level. Longitudinal studies have shown significant increases in DIT scores for individuals in their college years. There is also evidence that specific interventions (e.g. ethics training programs) can improve a subject's DIT score. However, changes are expected to be a long-term process and not the result of, for example, short exposure to a single game.

3.2.2 Intuitionism

Moral Foundations Theory (MFT) is based on the Social Intuitionist model (SIM) of morality championed by Haidt (2001) and others. This model (rooted in a Humean tradition) stands in opposition to Kohlberg's rationalist model (rooted in a Kantian tradition). It claims that moral judgment is primarily intuitive and driven more by emotion than by reasoning. Based on this model, Haidt and his colleagues performed a large-scale study to identify the intuitive foundations of morality across many societies. They identified five different moral foundations: Care/harm, Fairness/cheating, Loyalty/betrayal, Authority/subversion, and Sanctity/degradation. Another foundation, Liberty/oppression, has since been proposed (Iyer et al. 2012). Studies have shown that these foundations are widely recognised across cultures, but that different cultures and political groups place different emphases on each. For example, US liberals place greater importance on the "individualising" foundations of Care and Fairness, and less on the "binding" foundations of Loyalty, Authority and Sanctity, while conservatives place importance more evenly across all five (Graham et al. 2009). However, further research challenges the claim that there are only five foundations, and this has recently resulted in the development of alternative instruments (Curry et al. 2019). Unlike the schemas of the DIT, there is no claim that different moral foundations represent higher or lower levels of moral development. The SIM is not a developmental theory and expressly avoids making any prescriptive claims. We describe three evaluation tasks based on the MFT. The first, the MFQ, is by far the most prominent.

Moral Foundations Questionnaire (MFQ)

The MFQ is a test to determine how strongly subjects rate each of the five primary moral foundations (Graham et al. 2011). The test contains two parts. In the first, subjects are asked "*When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?*" Sixteen different criteria are presented, based on the different foundations, and the subject is asked to rate their importance. In the second part the subject is asked to agree or disagree on a Likert scale with moral statements such as "*It can never be right to kill a human being.*" The test produces a score for each of the five foundations showing its level of importance to the subject.

The MFQ has been thoroughly validated, however its internal consistency is not high, indicating a large variability in answers to the different questions pertaining to each foundation (Graham et al. 2011). The MFT does not distinguish whether responses are generated intuitively or deliberately. There is some evidence that placing subjects under cognitive load while taking the test can influence their choices (Wright and Baril 2011), which could indicate that deliberation is affecting answers in normal circumstances.

Moral Foundations Vignettes (MFV)

The MFV is a test of moral focus (Clifford et al. 2015). It consists of 132 brief micro-moral scenarios, such as "*You see a teenage boy chuckling at an amputee he passes by while on the subway*", which are intended to relate to the six moral foundations. Subjects are asked to rate the wrongness of each scenario on a Likert scale. The test has been validated; however, it has not been widely adopted yet.

Moral Foundations Sacredness Scale (MFSS)

The MFSS measures the subject's self-expressed willingness to perform potentially immoral acts for money in order to assess "the degree to which people sacralise each of the five foundations" (Graham and Haidt 2012). It describes 24 micro-moral actions, such as "*Kick a dog in the head, hard*", drawn from the five main moral foundations, and asks the subject to state how much they would need to be paid to perform the action, from "*I'd do it for free!*", incrementally through to $1 million and finally "*never for any amount of money*". As a measure of sacredness, the test looks for those items which the subject would refuse to do at any price. The MFSS has been extensively validated but has low internal consistency.

Guidelines for use: The MFQ and its variants (MFV and MFSS) measure which foundations are most intuitively important for subjects when they evaluate moral actions, but they do not, unlike the various cognitive developmentalist tests, ask the subject to make and justify moral decisions about what action should be done. This suggests that it is a test for moral focus, insofar as subject's rank what is most morally important to them, although it also has judgment elements. Unlike the work of Kohlberg and Rest, the SIM is not a developmental model and doesn't make any claims about one foundation being higher or better than another. As such, there is little work exploring how a person's moral foundations change over time. The MFQ should thus not be used as a measure for long-term moral development. The MFQ is comparatively simple to administer and easy to score.

3.2.3 Prosocial Morality

The moral psychology of Kohlberg and Rest, which underlies the MJI, SRM and DIT tests, has been criticised as gender-biased, favouring a 'masculine' ethics of justice over a 'feminine' ethics of care (Gilligan 1982). While the validity of this as a gender-specific difference has been called into serious question (Lapsley 1996, p. 134), Gilligan's critique of cognitive developmentalism brought to the fore the latter's (acknowledged) focus on justice as the sole basis for mature moral judgment, excluding more altruistic

and “care” oriented concerns. In response to this, Eisenberg (1986) developed a model of prosocial moral reasoning, focusing on conflict between a person’s own needs and desires and those of others. She proposes six distinct schemas applied to prosocial reasoning, “Hedonistic”, “Needs-oriented”, “Approval-oriented”, “Stereotypic”, “Sympathetic”, and “Internalised affect”.

Measure of Prosocial Moral Reasoning (PROM)

The PROM (Carlo et al. 1992) is an evaluation task, similar in structure to the DIT, for measuring the subject’s preference for each schema. Longitudinal studies have shown PROM scores are consistent from adolescence to early adulthood, and correlate with other measures of prosocial behaviour (Eisenberg et al. 2002).

Measure of Moral Orientation (MMO)

The MMO is designed to measure a subject’s preference for either justice- or care-oriented ethics (Liddell and Davis 1992). It comprises two components: a judgment and a self-description component, each measured on a Likert scale. The test is aimed at US college students, and the dilemmas are micro-moral situations based on their everyday lives. It is scored in four parts with separate care and justice scores for both the judgment and self-description components. It has been validated, but not extensively (Liddell and Davis 1992).

Guidelines for use: Prosocial morality focuses on acts that are morally good, but not strictly required by justice. As a test of moral judgment, the PROM is more suitable for evaluating a subject’s altruistic and caring tendencies, while the DIT and associated tests focus on what is morally permissible or impermissible. The kind of moral problems being studied will determine which test is more appropriate. In instances where aspects of both are in play, the MMO can be used to reveal subject’s preferences for either justice or care ethics. The MMO’s use of micro-moral dilemmas sets it apart from the macro-morality of the previously described tests, but also makes it more culturally specific, limiting its use.

3.2.4 Moral Identity

Moral identity research locates morality as a key element in the construction of our self-image and social identity (Aquino and Reed 2002; Blasi 1993).

Measure of Moral Identity (MMI)

Aquino and Reed (2002) propose the MMI as an instrument to measure the overall importance of moral traits as part of a subject’s self-identity. The test names nine morally relevant traits (*caring*, *compassionate*, *fair*, *friendly*, *generous*, *helpful*, *hardworking*, *honest*, and *kind*) and asks the subject to visualise a person with these characteristics, then agree or disagree (on a Likert scale) with statements describing how much they desire to be like that person, and how often they behave in ways that demonstrate those traits. These rankings are used to calculate two scores, for *Internalisation* and *Symbolism*, rating their inward and outward expression of morality as part of their self-identity.

Triune Ethics Orientation(TEO)

Narvaez (2016) expands the idea of moral identity to recognise that there are several moral mindsets that can drive our self-image. Her Triune Ethics Metatheory recognises three distinctive ethical orientations: *Protectionist*, *Engagement* and *Reflective*. The TEO (Narvaez and Hardy 2016) is a measure of the relative importance of these three orientations to an individual. For each orientation, the subject is shown a set of related words and asked to agree or disagree with statements describing their self-perception of these characteristics. These responses are used to calculate scores for each orientation.

Guidelines for use: Moral identity is a key element of our moral focus. We should expect, for instance, that players with a strong moral identity would place more importance on moral choices in a game. The TEO could be used to further refine this, depending on how the moral situations and possible responses align with different moral mindsets.

3.2.5 Value Theory

Value theory is a study of what motivates people, with "values" defined as desirable goals that transcend specific situations and serve as standards to guide action. Values form a hierarchical system and can conflict with one another (Schwartz 2012). Schwartz (1992) identifies ten broad values: Self-direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence and Universalism. These ten values can be simplified to positions on a continuum with two orthogonal axes: openness to change vs conservation, and self-enhancement vs self-transcendence. The two instruments below have been developed to measure the relative importance subjects give to each of the ten values.

Schwartz Values Scale (SVS)

The SVS presents a list of 56 "value items", each of which expresses a motivational goal of a particular value. Subjects rate each one on a 9-point scale from -1 (*opposed to my values*), 0 (*not important*), to 7 (*of supreme importance*). Scores for each of the ten values are calculated by averaging the scores for each associated item. A cross-cultural study of over 20,000 subjects in 40 different countries shows that the test reliably measures the ten value types and demonstrates the universality of the two orthogonal dimensions (Schwartz and Sagiv 1995).

Portrait Values Questionnaire (PVQ)

The PVQ is an alternative to the SVS featuring simpler language, making it easier to administer to younger or less educated subjects (Bubeck and Bilsky 2004). It avoids the abstract concepts presented in the SVS in favour of concrete verbal portraits. Each portrait describes a person's aspirations, pointing towards a particular value, with 40 items in all. The score for each value is the average of the scores for each related item. The PVQ has been thoroughly cross-culturally tested with N=35,161 subjects in 20 different countries, demonstrated to be a reliable measure, and a strong predictor of cultural values.

Guidelines for use: Both the SVS and the PVQ might be understood as (indirect) measures of moral focus, as they measure how a subject ranks moral motivations, such as Benevolence and Universalism, against their self-

interested pursuit of Power, Achievement and Hedonism, although there is often a gap between how subjects complete evaluation tasks and how they act in practice.

3.2.6 Other Tests

The remaining tests are either not as well established, as relevant to game researchers, or as widely used as those discussed above, but are worth mentioning because they investigate aspects of moral psychology not captured by the above measures. For that reason, they will be described more briefly.

The Ethics Position Questionnaire (EPQ)

The EPQ attempts to classify the differences in the way individuals arrive at moral judgments (Forsyth 1980). Forsyth posits that there are two factors at play: 1) *Universalism/Relativism*: the degree to which a person believes in universal moral standards; 2) *Idealism/Realism*: the degree to which a person believes moral ideals can always be achieved. The EPQ attempts to assign subjects to four categories by scoring them on the scales of Universalism/Relativism and Idealism/Realism. It consists of 20 position statements, which subjects are asked to agree or disagree with on a Likert scale. Forsyth found that there was no significant correlation between scores on the EPQ and DIT, which indicates that ethical ideology is a distinct construct to the DIT's measure of moral development.

Guidelines for use: The EPQ can be considered a test of moral judgment by characterising the broad philosophical stances with which subjects reach moral conclusions.

Measure of Ethical Viewpoints (MEV)

Brady and Wheeler (1996) propose the MEV as a measure of a subject's preference for formalist (i.e. based on social norms) versus consequentialist (i.e. based on perceived harm) ethical principles in making judgments. It consists of two sections. The first section is structured similarly to the DIT, consisting of eight vignettes each presenting a micro-moral dilemma. For each vignette, two formalist and two consequentialist rationales are presented, one in favour of the action and one against. The second part of the test focuses on twenty character traits associated with one of the two predispositions (e.g. "*principled*", "*resourceful*").

Guidelines for use: The MEV is a measure of moral judgment (first part) and focus (second part). Similarly to the EPQ, it characterises the broad philosophical stance with which a subject reaches moral judgments.

The Moralization of Everyday Life (MELS)

Lovett et al. (2012) criticise the widespread use of unnatural life-or-death hypotheticals in the study of moral judgment. They designed the MELS as an alternative measure which focuses on "moral judgments of everyday behaviours". The scale contains descriptions of 30 actions (such as "*Keeping extra money accidentally dispensed from an ATM*") which subjects are asked to rate on a scale from not wrong at all to an extremely immoral action. It is scored

based on six factors, including lying, harm, laziness, acting beyond duty, body violations and disgusting behaviours.

Guidelines for use: The MELS is a test of moral judgment in everyday micro-moral situations. Its applicability, however, is limited due to the specific social situations it describes.

Multiple Intelligence Profiling Questionnaire – Ethical Sensitivity Scale (MIPQ-ESS)

The MIPQ (Tirri and Nokelainen 2011) is a collection of instruments for measuring subject's ability to solve problems, based on Gardiner's theory of multiple "intelligences". Included among these is an Ethical Sensitivity Scale (ESS) which measures a subject's self-evaluation of their ability in each of the seven skill areas (such as reading emotions and perspective taking) identified by Narvaez and Endicott (2009) as necessary for ethical sensitivity. The subject is presented with 28 self-assessment statements, four associated with each skill area, and asked to indicate whether they agree with each on a Likert scale.

Guidelines for use: While this instrument is a measure of moral sensitivity, it is based, not on a performance test, but on the subject's self-assessment of their ability.

Immediate Affect Toward Moral Stimuli (IAMS)

The Affect Misattribution Procedure described earlier has been employed by Hofmann and Baumert (2010) as a measure of moral sensitivity. They used images depicting both positive and negative moral behaviours as priming stimuli, as well as positive and negative non-moral images. Subjects were asked to rate the target stimuli (Chinese pictographs) on a scale from "very positive" to "very negative". An IAMS score was calculated by the proportion of positive judgments based on a positive moral prime minus the number of positive judgments based on a negative moral prime. This score was shown to correlate well with moral emotions experienced in later tests. Subjects with a high IAMS score (indicating high moral sensitivity) felt more guilt over the outcome of a hypothetical trolley problem, and more anger at an unfair outcome in an ultimatum game.

Guidelines for use: As a measure of intuitive moral sensitivity, the IAMS can work as a predictor of moral behaviour in games, particularly when players are under pressure and do not have time to deliberate about their behaviour.

Related Tests In Brief

There are many more psychometric tests that measure aspects of psychology that affect moral cognition, such as:

- Empathy (Baron-Cohen and Wheelwright 2004; Davis 1983)
- Disgust (Schnall et al. 2008)
- Dark Triad of Machiavellianism, narcissism, and psychopathy (Jones and Paulhus 2014)
- Guilt and Shame (Cohen et al. 2011; Kugler and Jones 1992)

Guidelines for use: While, for reasons of space, we shall not explore these in detail here, they can all be important for different research projects. For example, using a Guilt and Shame scale to measure how guilty a player feels

after making an in-game choice that conflicts with their real-world moral values can tell us something about their level of moral engagement (Weaver and Lewis 2012). We might also, for example, use an empathy scale, such as the Interpersonal Reactivity Index (Davis 1983), to determine if players with higher empathy scores are more likely to favour deontological moral choices over utilitarian ones in gameplay.

3.3 Response Measures

As well as recording the moral behaviours that players make during gameplay, including their ability to persist at moral action in the face of adversity, we can also use response dynamics to examine whether Type 1 or Type 2 processes are involved in in-game moral decisions. The most sophisticated experiments use fMRI imaging to detect activation of particular regions of the brain (Greene et al. 2001), but other simpler and less intrusive measures can also be employed, such as measuring response times and mouse movement.

Response Time

If dual-process theory is correct, we should expect to see longer response times when deliberation is required to make a decision, while intuitive choices should be made more quickly. Moore et al (2008) measured subject's response times to a set of hypothetical moral scenarios involving the sacrifice of one life to save several others. Each dilemma was presented as either a "personal" version in which the subject kills the victim through physically direct action (e.g. pushing someone), and an "impersonal" version in which the death occurs at a greater psychological distance (e.g. throwing a switch). Response times were significantly greater for the impersonal variations, which is taken as evidence that personal dilemmas invoke emotional (intuitive) responses while impersonal dilemmas involve deliberative reasoning. Similarly, Suter & Herwig (2011) showed that limiting a subject's time to judge a personal moral dilemma resulted in more intuitive responses. However, while this effect has been the subject of several studies (Baron and Gürçay 2017) and is central to the discussion of utilitarian and deontological moral theories, it remains contested (McGuire et al. 2009).

Mouse Movement

Koop and Johnson (2013) propose a technique for exposing subjects' decision processes by examining mouse trajectories. In experiments they presented two alternatives as buttons in the top left and right corners of a computer screen. With the mouse cursor starting at the bottom middle of the screen, they recorded its trajectory as the subject selected their preference. When the preferred alternative was obvious, the cursor moved in a fairly direct line to the button, but the closer the comparison, the more curved the path was, indicating a "competitive pull" towards the alternative. In a more complex case, choosing between safer or riskier gambles, a more distinctive path was evident, moving initially towards the safe option before switching to the riskier alternative. This is taken as evidence of a "default" intuitive choice being overridden by a deliberate decision. Koop (2013) has since extended this work to moral decision-making, with the expectation of seeing a similar dual-process effect in choosing between intuitive "deontological" and deliberate "utilitarian"

responses. Contrary to expectations, the trajectories gave no evidence of dual-process decision-making. Nonetheless, the technique warrants attention for the insight it provides into decision-making. Eye tracking data might be similarly useful (Fiedler et al. 2013).

Guidelines for use: Measuring mouse movement might be a way to infer the type of cognitive processes that are underwriting moral behaviour during gameplay. Placing players under cognitive load or time pressure while asking them to make moral judgments in games should lead to more intuitive, deontological judgments, especially in personal moral dilemmas. Choices made quickly may also indicate an intuitive Type 1 response, whereas slower choices may indicate a deliberative Type 2 response.

4. MORALITY MEASURES IN GAMES RESEARCH

To help assist games researchers with selecting an appropriate instrument for their research, we look at some existing research examples below. But first we present a series of questions they should ask themselves:

1. What is the causal relationship between morality and gameplay you wish to explore? Are you investigating *moral engagement* (how morality impacts gameplay) or *moral effects* (how gameplay impacts morality)?

2. What theoretical model are you following? Are you looking to investigate Type 1 or Type 2 processes, or the interaction between the two? Note, different tests carry with them a commitment to different theoretical models of morality.

3. Which component of moral engagement are you examining: focus, sensitivity, judgment, or action?

4. What is the most appropriate instrument that suits your needs? Has it been validated and shown to be a reliable measure? How hard is it to administer?

5. What in-game behaviours should you measure? This includes in-game choices and actions, and the dynamics of their responses, including speed, mouse movement and eye-tracking data.

For example, an experiment to establish whether player's real-world moral motivations carry over to in-game choices might use the MFQ to establish the players' implicit moral foundations and compare these to the in-game choices they make. However, in doing so you should be aware that the MFQ is intended as a measure of Type 1 moral processes and evidence shows that it can be influenced by time-pressure or cognitive load (Wright and Baril 2011).

Alternatively, we might be interested in factors that affect a player's moral judgments in games by comparing utilitarian and deontological moral reasoning. As a pre-test we might use the MEV or EPQ to establish the player's preference for consequentialist or formalist moral reasoning. Then in the game we could present moral dilemmas that pit utilitarian against deontological judgments, both under time pressure and not. Response times and mouse

movement dynamics could provide insight into the underlying mental processes involved.

Another question might address our sensitivity to moral themes in games. Often moral choices in games are clearly signposted and their significance laid bare for the player, however in more system-driven games the moral themes may be more implicit (Formosa et al. 2016). Evidence suggests that our ability to detect and understand moral themes embedded in narratives depends on the presence of appropriate moral schemas (Narvaez 2001). For example, pre-conventional thinkers tend to extract pre-conventional themes from moral scenarios, irrespective of authorial intent. Using the DIT as a pre-test measure, we might see whether the same holds true of interactive narratives, matching in-game decisions to DIT scores and post-test self-report responses.

The various instruments can also be employed to help investigate short- and long-term moral effects. One option is to see if playing a game has short-term impacts on various morality measures. For example, compared to a control group, does playing a game about refugees change the importance of Benevolence in the SVS or the Care/Harm foundation in the MFQ? And if it does, how long does that effect last? Since one-off short sessions of play should not be expected to have long-term impacts, to measure such long-term effects longitudinal field studies of real-world players, rather than short laboratory studies, seem most appropriate. The various instruments outlined here might be used to measure if any moral shifts occur after a long time spent playing a particular game or type of game (such as pro-social or violent games).

4.1. Existing research

While empirical research into morality in videogames is in a comparatively nascent state, there are a few recent studies that illustrate how this research can be successfully undertaken. Much of this research uses the MFQ to explore moral engagement. Krcmar and Cingel (2016) use the MFQ as a pre-test to establish a player's moral foundations, which are then compared with their in-game motivations expressed using a think-aloud protocol. Transcripts of the players' expressed reasoning were coded as strategic or moral, with moral reasons coded according to particular moral foundations. Significant correlations between scores on the MFQ pre-test and in-game moral reasoning were found for the Fairness/Cheating and Authority/Subversion foundations, but not for the other foundations. However, caution must be applied when using verbal transcripts, since these may represent post-facto rationalisations. A separate study by Joeckel et al. (2012) also found that pre-test MFQ scores were significantly correlated with in-game decisions. Weaver and Lewis (2012) analyse the different choice scenarios in *Fallout 3* in terms of the five moral foundations. They recorded the in-game choices subjects made while playing this game and demonstrated that the pre-test responses subjects gave in the MFQ predicted their in-game decisions. They backed this assessment up with the use of post-test enjoyment and guilt scales, which showed that players felt guilty after behaving in anti-social ways in the game, although this did not impact on their enjoyment. The MFQ has also been used by Grizzard et al. (2014) as a post-test to demonstrate that playing a guilt-inducing game leads to higher scores for relevant moral foundations. The authors also show that playing the

guilt-inducing game led to higher scores for the Care/Harm and Fairness/Cheating foundations when compared to simply asking participants to recall a time they felt guilty.

Focusing more on moral effects, Narvaez and Mattan (2008) propose a prosocial alternative to the much-debated General Aggression Model (GAM) (Dewall, et al. 2011). If we assume, with the GAM, that violent videogames activate/reinforce hostile and aggressive schemas, then prosocial videogames ought to activate/reinforce prosocial schemas. To test this hypothesis, the authors use a mod made for Bioware's *Neverwinter Nights* (2002) in which subjects are tasked with slaying bandits (violent condition), helping the sick (prosocial condition), or collecting bags of gold (neutral condition). Participants were then asked to complete a series of "story stems" which were scored for aggressive, neutral, and prosocial content. As hypothesised, the helping version led to significant increases in prosocial responses, with those who played the game more likely to describe story characters in empathetic terms. The violent version of the game, however, did not lead to a significant increase in violent or aggressive responses, with aggressive responses remaining more or less equal across all conditions (including several controls). Similarly, alternative techniques used in studying other forms of media, such as spontaneous trait inferencing and lexical decision, might also be applied to videogames research (e.g. Narvaez et al. 2006).

These examples illustrate some of the ways in which the instruments of moral psychology have been successfully applied in games research. To date, the MFQ seems to have been used more widely in games research than the other available instruments, perhaps because it is simple to administer and score. The use of other well-established morality measures in games research has so far been limited, which could be restricting progress in the field. We hope this paper will help inspire future innovations in this area and the use of a wider range of relevant instruments.

5. CONCLUSION

The extensive existing and ongoing research in moral psychology leaves those of us in games research with a responsibility and an opportunity. Our responsibility is to do rigorous research by understanding the theoretical commitments we are making in our work and employing methods that are theoretically appropriate and proven to be valid and reliable. Our opportunity comes from the wealth of researchers who have gone before us and established a variety of methods and instruments for investigating moral decision-making that can be applied, with care, to games research. Research in games and ethics suffers from an abundance of theory and a lack of empirical results. To address this deficiency, we need to employ a set of standardised and well-validated instruments that measure the aspects of moral cognition players use to play such games. We hope that this review helps games researchers find the tools that are appropriate to their goals.

BIBLIOGRAPHY

Agle, B. et al. 2014. *Research Companion to Ethical Behavior in Organizations*.

Cheltenham: Edward Elgar.

Anderson, C. et al. 2010. “Violent Video Game Effects on Aggression, Empathy, and Prosocial Behavior in Eastern and Western Countries.” *Psychological Bulletin* 136(2):151–73.

Aquino, K., Reed, A. 2002. “The Self-Importance of Moral Identity.” *Journal of Personality and Social Psychology* 83(6):1423–40.

Bargh, J., Chartrand, T. 2000. “The Mind in the Middle.” Pp. 253–85 in *Handbook of research methods in social and personality psychology*, Ed. Reis and Judd. Cambridge: Cambridge UP.

Baron, J., Gürçay, B. 2017. “A meta-analysis of response-time tests of the sequential two-systems model of moral judgment.” *Memory and Cognition* 45(4):566-75.

Baron-Cohen, S., Wheelwright, S. 2004. “The Empathy Quotient.” *Journal of Autism and Developmental Disorders* 34(2):163–75.

Basinger, K., Gibbs, J., Fuller, D. 1995. “Context and the Measurement of Moral Judgment.” *International Journal of Behavioural Development* 18(3):537–56.

Belman, J., Flanagan, M. 2010. “Designing Games to Foster Empathy.” *International Journal of Cognitive Technology* 15(1):5–15.

Bioware. 2002. *Neverwinter Nights* [PC game], Infogrames

Blasi, A. 1980. “Bridging moral cognition and moral action.” *Psychological Bulletin* 88(1):1-45.

Blasi, A. 1993. “The Development of Identity.” Pp. 99-122 in *The Moral Self,* Ed. Noam and Wren. Cambridge: MIT Press.

Brady, F., Wheeler., G. 1996. “An Empirical Study of Ethical Predispositions.” *Journal of Business Ethics* 15(9):927–40.

Bubeck, M., Bilsky, W. 2004. “Value structure at an early age.” *Swiss Journal of Psychology* 61(1):31–41.

Cameron, D., et al. 2013. “Morality in High Definition.” *Journal of Experimental Social Psychology* 49(4):719–25.

Carlo, G., Eisenberg, N., Knight., G. 1992. “An Objective Measure of Adolescents’ Prosocial Morality.” *Journal of Research on Adolescence* 2(4):331–49.

Clifford, S., et al. 2015. “Moral Foundations Vignettes.” *Behavior Research Methods* 47(4):1178–98.

Cohen, T. et al. 2011. “Introducing the GASP Scale.” *Journal of Personality and Social Psychology* 100(5):947–66.

Colby, A. et al. 1987. *The Measurement of Moral Judgment: Volume 2*. Cambridge: Cambridge UP.

Colby, A. and Kohlberg, L. 1987. *The Measurement of Moral Judgment, Volume 1*. Cambridge: Cambridge UP.

Curry, O., Chesters, M., van Lissa, C. 2019. “Mapping morality without a

compass." *Journal of Research in Personality* 78:106–24.

Davis, M. 1983. "Measuring Individual Differences in Empathy." *Journal of Personality and Social Psychology* 44(1):113–26.

DeWall, C., Anderson, C., Bushman, B. 2011. "The general aggression model." *Psychology of Violence* 1(3): 245-58.

Eisenberg, N. 1986. *Altruistic Emotion, Cognition, and Behavior*. Hillsdale: Lawrence Erlbaum.

Eisenberg, N. et al. 2002. "Prosocial Development in Early Adulthood." *Journal of Personality and Social Psychology* 82(6):993–1006.

Evans, J. and Stanovich, K. 2013. "Dual-process theories of higher cognition" *Perspectives on psychological science*, 8(3): 223-41.

Ferguson, C. 2007. "The Good, the Bad and the Ugly." *Psychiatric Quarterly* 78(4):309–16.

Fiedler, Susann, et al. 2013. "Social Value Orientation and Information Search in Social Dilemmas." *Organizational Behavior and Human Decision Processes* 120(2):272–84.

Flanagan, Mary, et al. 2007. "A Method for Discovering Values in Digital Games." *Proceedings of DiGRA 2007 Conference*.

Formosa, P, Ryan, M., Staines, D. 2016. "Papers, Please and the systemic approach to engaging ethical expertise in videogames." *Ethics and Information Technology* 18(3): 211-25.

Forsyth, D. 1980. "A Taxonomy of Ethical Ideologies." *Journal of Personality and Social Psychology* 39(1):175–84.

Gibbs, J., et al. 2007. "Moral Judgment Development across Cultures." *Developmental Review* 27(4):443–500.

Gibbs, J., Widaman, K., Colby, A. 1982. "Construction and Validation of a Simplified, Group-Administerable Equivalent to the Moral Judgment Interview." *Child Development* 53(4):895–910.

Gilligan, C. 1982. *In a Different Voice*. Cambridge: Harvard UP.

Graham, J. et al. 2011. "Mapping the Moral Domain." *Journal of Personality and Social Psychology* 101(2):366–85.

Graham, J. Haidt, J. 2012. "Sacred Values and Evil Adversaries." Pp. 11–31 in *The social psychology of morality*, Ed. Mikulincer and Shaver. American Psychological Association.

Graham, J., Haidt, J. Nosek, B. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology* 96(5):1029–46.

Greene, J., et al. 2008. "Cognitive Load Selectively Interferes with Utilitarian Moral Judgment." *Cognition* 107(3):1144–54.

Greene, J., et al. 2001. "An FMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293(5537):2105–8.

Grizzard, M., et al. 2014. "Being Bad in a Video Game Can Make Us Morally

Sensitive." *Cyberpsychology, Behavior, and Social Networking* 17(8):499–504.

Haidt, J. 2001. "The Emotional Dog and Its Rational Tail." *Psychological Review* 108(4):814–34.

Haidt, J. Bjorklund, F. 2007. "Social Intuitionists Answer Six Questions about Morality." Pp. 181–218 in *Moral Psychology: Volume 2*, Ed. Sinnott-Armstrong. Cambridge: MIT Press.

Hoffman, W. Baumert, A. 2010. "Immediate Affect as a Basis for Intuitive Moral Judgment." *Cognition and Emotion* 24(3):522–35.

Iyer, R., et al. 2012. "Understanding Libertarian Morality." *PloS One* 7(8).

Joeckel, S., Bowman, N., Dogruel, L. 2012. "Gut or Game?" *Media Psychology* 15:460–85.

Jones, D., Paulhus D. 2014, "Introducing the Short Dark Triad (SD3)." *Assessment* 21(1): 28-41.

Jordan, J. 2007. "Taking the First Step Toward a Moral Action: A Review of Moral Sensitivity Measurement Across Domains." *The Journal of Genetic Psychology* 168(3):323–59.

Kohlberg, L. 1981. *Essays on Moral Development (Vol. 1)*. San Francisco: Harper & Row.

Koop, G. 2013. "An Assessment of the Temporal Dynamics of Moral Decisions." *Judgment and Decision Making* 8(5):527–39.

Koop, G., Johnson, J. 2013. "The Response Dynamics of Preferential Choice." *Cognitive Psychology* 67(4):151–85.

Krcmar, M., Cingel. D. 2016. "Moral Foundations Theory and Moral Reasoning in Video Game Play." *Journal of Broadcasting and Electronic Media* 60(1):87–103.

Krebs, D., Denton, K., Wark, G. 1997. "The Forms and Functions of Real-Life Moral Decision-Making." *Journal of Moral Education* 26(2):131–45.

Kugler, K., Jones, W. 1992. "On Conceptualizing and Assessing Guilt." *Journal of Personality and Social Psychology* 62(2):318–27.

Lapsley, D. 1996. *Moral Psychology*. Boulder: Westview Press.

Lapsley, D., Narvaez, D. 2005. "Moral Psychology at the Crossroads." Pp. 18-35 in *Character Psychology and Character Education,* Ed. Lapsley and Power. Notre Dame: University of Notre Dame Press.

Lapsley, D., Hill, P. 2008. "On Dual Processing and Heuristic Approaches to Moral Cognition." *Journal of Moral Education* 37(3):313–32.

Liddell, D., Davis, T. 1992. "The Measure of Moral Orientation." *Journal of College Student Development* 37(5):485–93.

Lind, G. 2013. "30 Years of the Moral Judgment Test." Pp. 143–170 in *Estudos e pesquisas em psicologia do desenvolvimento e da personalidade,* Ed. Hutz and de Souza. Sao Paulo: Casa do Psicólogo.

Lind, G. 2019. "Using the Moral Competence Test (MCT)" *Moral Democratic*

Competence, https://www.uni-konstanz.de/ag-moral/mut/mjt-engl.htm

Lovett, B., Jordan, A., Wiltermuth, S. 2012. "Individual Differences in the Moralization of Everyday Life." *Ethics and Behavior* 22(4):248–57.

McGuire, J., et al. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577–80.

MicroProse. 1991. *Civilization* [PC Game], MicroProse.

Molleindustria. 2006. *McDonalds Videogame* [Browser Game], Molleindustria

Moore, A., Clark, B., Kane, M. 2008. "Who Shalt Not Kill ?" *Psychological Science* 19(6):549–57.

Narvaez, D. 2001. "Moral text comprehension." *Journal of Moral Education* 30(1): 43-54.

Narvaez, D. 2016. *Embodied Morality*. New York: Palgrave Macmillan.

Narvaez, D., et al. 2006. "Moral chronicity and social information processing." *Journal of Research in Personality* 40:966-85.

Narvaez, D., Bock, T. 2002. "Moral Schemas and Tacit Judgment or How the Defining Issues Test Is Supported by Cognitive Science." *Journal of Moral Education* 31(3):297–314.

Narvaez, D., Endicott, L. 2009. "Ethical Sensitivity." in *Nurturing Character in the Classroom, EthEx series*. Notre Dame: University of Notre Dame Press.

Narvaez, D., Hardy, S. 2016. "Measuring Triune Ethics Orientations." Pp. 47–72 in *Embodied Morality*. New York: Palgrave Macmillan.

Narvaez, D., Mattan, B. 2008. "Kill bandits, collect gold or save the dying." Media Psychology Review 1(1).

Nederhof, A. 1985. "Methods of Coping with Social Desirability Bias." *European Journal of Social Psychology* 15(3):263–80.

Payne, K., Lundberg, K. 2014. "The Affect Misattribution Procedure." *Social and Personality Psychology Compass* 8(12):672–86.

Pennycook, G., Fugelsang, J., Koehler, D. 2015. "What Makes Us Think?" *Cognitive Psychology* 80:34–72.

Polanyi, M. 1966. *The Tacit Dimension*. Chicago: Chicago University Press.

Rest, J. 1979. *Revised Manual for the Defining Issues Test*. Minneapolis: Minnesota Moral Research Projects.

Rest, J. 1983. "Morality." pp. 556–629 in *Handbook of child psychology: Vol 3.* Ed. Flavell and Markham. Hillsdale: Lawrence Erlbaum.

Rest, J., et al. 1997A. "Designing and Validating a Measure of Moral Judgment." *Journal of Educational Psychology* 89(1):5–28.

Rest, J., et al. 1997B. "Alechemy and Beyond: Indexing the Defining Issues Test." *Journal of Educational Psychology* 89(3):498–507.

Rest, J., et al. 1999A "A neo-Kohlbergian approach." *Educational Psychology*

Review 11(4): 291-324.

Rest, J., et al. 1999B. "DIT2." *Journal of Educational Psychology* 91(4):644–59.

Rest, J., et al. 1999C. *Postconventional moral thinking*. Mahwah: Lawrence Erlbaum.

Ryan, M., Staines, D., Formosa, P. 2017. "Focus, Sensitivity, Judgment, Action." *Transactions of the Digital Games Research Association* 3(2).

Samuels, P., 2015. "Advice on Reliability Analysis with Small Samples". DOI: 10.13140/RG.2.1.1495.5364.

Schnall, S., et al. 2008. "Disgust as Embodied Moral Judgment." *Personality and Social Psychology Bulletin* 34(8):1096–1109.

Schrier, K. 2015. "EPIC." *Journal of Moral Education* 44(4):393–424.

Schwartz, S. 2012. "An Overview of the Schwartz Theory of Basic Values." *Online Readings in Psychology and Culture* 2(1).

Schwartz, S. 1992. "Universals in the Content and Structure of Values." *Advances in experimental social psychology* 25:1-65.

Schwartz, S., Sagiv, L. 1995. "Identifying Culture-Specifics in the Content and Structure of Values." *Journal of Cross-Cultural Psychology* 26(1):92–116.

Sicart, M. 2013. *Beyond Choices*. Cambridge: MIT Press.

Suter, R., Hertwig, R. 2011. "Time and Moral Judgment." *Cognition* 119(3):454–58.

Telltale Games 2012. *The Walking Dead* [PC Game], Telltale Games.

Thoma, S., et al. 2013. "Describing and testing an intermediate concept measure of adolescent moral thinking." *European Journal of Development Psychology* 10(2):239-52.

Thoma, S. 2014. "Measuring Moral Thinking from a Neo-Kohlbergian Perspective." *Theory and Research in Education* 12(3):347–65.

Tirri, K. Nokelainen, P. 2011. *Measuring Multiple Intelligences and Moral Sensitivities in Education*. Rotterdam: Sense Publishers.

Vozzola, E. 2014. *Moral Development*. New York: Routledge

Weaver, A. Lewis, N. 2012. "Mirrored Morality." *Cyberpsychology, Behavior, and Social Networking* 15(11):610–14.

Wright, J., Baril, G. 2011. "The Role of Cognitive Resources in Determining Our Moral Intuitions." *Journal of Experimental Social Psychology* 47(5):1007–12.

Zagal, J. 2011. "Ethical Reasoning and Reflection as Supported by Single-Player Videogames." Pp. 19–35 in *Designing games for ethics*, Ed. Schrier and Gibson. Hershey: IGI Global.