Journal of Educational Psychology
1999, Vol. 91, No. 4, 644–659

0022-0663/99/$3.00

DIT2: Devising and Testing a Revised Instrument of Moral Judgment

James R. Rest and Darcia Narvaez
University of Minnesota, Twin Cities Campus

Stephen J. Thoma
University of Alabama, and University of Minnesota, Twin Cities Campus

Muriel J. Bebeau
University of Minnesota, Twin Cities Campus

The Defining Issues Test, Version 2 (DIT2), updates dilemmas and items, shortens the original Defining Issues Test (DIT1) of moral judgment, and purges fewer participants for doubtful response reliability. DIT1 has been used for over 25 years. DIT2 makes 3 changes: in dilemmas and items, in the algorithm of indexing, and in the method of detecting unreliable participants. With all 3 changes, DIT2 is an improvement over DIT1. The validity criteria for DIT2 are (a) significant age and educational differences among 9th graders, high school graduates, college seniors, and students in graduate and professional schools; (b) prediction of views on public policy issues (e.g., abortion, religion in schools, rights of homosexuals, women's roles); (c) internal reliability; and (d) correlation with DIT1. However, the increased power of DIT2 over DIT1 is primarily due to the new methods of analysis (a new index called N2, new checks) rather than to changes in dilemmas, items, or instructions. Although DIT2 presents updated dilemmas and smoother wording in a shorter test (practical improvements), the improvements in analyses account for the validity improvements.

The Defining Issues Test, Version 2 (DIT2), is a revision of the original Defining Issues Test (DIT1), which was first published in 1974. DIT2 updates the dilemmas and items, shortens the test, and has clearer instructions. This is the third in a series of articles in the *Journal of Educational Psychology* aimed at improving the measurement of moral judgment (Rest, Thoma, & Edwards, 1997; Rest, Thoma, Narvaez, & Bebeau, 1997). Rest, Thoma, and Edwards (1997) proposed an operational definition of construct validity (seven criteria) that could be used to evaluate various measurement devices of moral judgment. Rest, Thoma, Narvaez, et al. (1997) reported that a new way of indexing DIT data, the N2 index, had superior performance on the seven criteria in contrast to the traditional P index, which has been used for over 25 years. (See the Rest, Thoma, Narvaez, et al., 1997, article for further discussion of N2.) This article reports on a revised version (new dilemmas and items) of the DIT1—the DIT2—with more streamlined instructions and shorter length. Also, this article describes a new approach to detecting bogus data ("new checks").

While we were reexamining aspects of the DIT1, we also reconsidered our methods of checking for participant reliability. That is, given a multiple-choice test that can be group administered—often under conditions of anonymity—some participants might fill out the DIT1 answer sheet without regard to test instructions, and some participants might give bogus data. The participant reliability checks are methods for detecting bogus data. For the past decades, we have used a procedure called "standard checks" to check for bogus data. In sum, DIT2 uses new checks instead of standard checks and uses revised items and dilemmas as well as N2. With these three changes, we wanted to see whether the research dividends would increase by creating alternatives to DIT1, P index, and standard checks.

However, we had an important question to consider before getting into the matter of updating: Why would anyone want a DIT score, either updated or not? Two issues are at the heart of the matter. First, is Kohlberg's approach so flawed that research ought to start anew? Second, can a multiple-choice test like the DIT (as opposed to interview data) yield useful information?

The Kohlbergian Approach

The DIT is derived from Kohlberg's (1976, 1984) approach to morality. In the past decades, many challenges to this approach have been made. Critics raise both philosophi-

James R. Rest, Department of Educational Psychology and Center for the Study of Ethical Development, University of Minnesota, Twin Cities Campus; Darcia Narvaez, College of Education and Human Development and Center for the Study of Ethical Development, University of Minnesota, Twin Cities Campus; Stephen J. Thoma, Department of Human Development, University of Alabama, and Center for the Study of Ethical Development, University of Minnesota, Twin Cities Campus; Muriel J. Bebeau, Department of Preventive Science and Center for the Study of Ethical Development, University of Minnesota, Twin Cities Campus. James R. Rest died in July 1999.

We thank Lee Fertig, Irene Getz, Carol Koskela, Christyan Mitchell, and Nanci Turner Shults for help in data collection. We also thank the Bloomington School District and the Moral Cognition Research Group at the University of Minnesota.

Correspondence concerning this article should be addressed to Darcia Narvaez, Department of Educational Psychology, University of Minnesota, 206 Burton Hall, 178 Pillsbury Drive Southeast, Minneapolis, Minnesota 55455. Electronic mail may be sent to narvaez@tc.umn.edu.

cal and psychological objections. In a recent book (Rest, Narvaez, Bebeau, & Thoma, 1999), the criticisms and challenges to Kohlberg's theory are reviewed and analyzed. In contrast to those who find Kohlberg's theory so faulty that they propose discarding it, we have found that continuing with many of Kohlberg's starting points has generated numerous findings in DIT research.

To appreciate how Kohlberg's basic ideas illuminate important aspects of morality, first consider a distinction between *macromorality* and *micromorality.* Just as in the field of economics, *macro* and *micro* distinguish different phenomena and different levels of abstraction in analysis, we use the terms to distinguish different phenomena and levels of analysis in morality. Macromorality concerns the formal structure of society, that is, its institutions, role structure, and laws. The following are the central questions of macromorality: Is this a fair institution (or role structure or general practice)? Is society organized in a way that different ethnic, religious, and subcultural groups can cooperate in it and should support it? Should I drop out of a corrupt society? On the other hand, micromorality focuses on the particular, face-to-face relationships of people in everyday life. The following questions are central to micromorality: Is this a good relationship? Is this a virtuous person? Both micro- and macromorality are concerned with establishing relationships and cooperation among people. However, micromorality relates people through personal relationships, whereas macromorality relates people through rules, role systems, and formal institutions. In macromorality, the virtues of acting impartially and abiding by generalizable principles are praised (for how else could strangers and competitors be organized in a societal system of cooperation?). In micromorality, the virtues of unswerving loyalty, dedicated care, and partiality are praised, because personal relationships depend on mutual caring and special regard. In our view, Kohlberg's theory is more pertinent to macromorality than to micromorality (for further discussion of macro- and micromorality, see Rest et al., 1999). Some of Kohlberg's critics fault his approach for not illuminating "everyday" morality (in the sense of micromorality; see Killen & Hart, 1995). However, it remains to be seen how well other approaches accomplish this.

The issues of macromorality are real and important, regardless of the relative contributions of a Kohlbergian or non-Kohlbergian approach to issues of micromorality. Regarding the importance of macromorality issues, consider Marty and Appleby's (e.g., 1991) six-volume series on current ideological clashes in the world. Marty and Appleby talked about the world's major disputes since the cessation of the Cold War. Formerly, the Soviet Union and Marxism/communism seemed to be the greatest threats to democracies. However, Marty and Appleby characterized the major ideological clash today as between fundamentalism and modernism; others describe the clash in ideology as the "culture war" between orthodoxy and progressivism (Hunter, 1991) or religious nationalism versus the secular state (Juergensmeyer, 1993). These clashes in ideology lead "to sectarian strife and violent ethnic particularisms, to skirmishes spilling over into border disputes, civil wars, and battles of secession" (Marty & Appleby, 1993, p. 1). Understanding how people come to hold opinions about macromoral issues is now no less important, urgent, and real than the study of micromoral issues. It is premature to say what approach best illuminates micromorality. However, we claim that a Kohlbergian approach illuminates macromorality issues (see Table 4.9 in Rest et al., 1999).

DIT1 research follows Kohlberg's approach in four basic ways. It (a) emphasizes cognition (in particular, the formation of concepts of how it is possible to organize cooperation among people on a society-wide scope); (b) promotes the self-construction of basic epistemological categories (e.g., reciprocity, rights, duty, justice, social order); (c) portrays change over time in terms of cognitive development (i.e., it is possible to talk of "advance" in which "higher is better"); and (d) characterizes the developmental change of adolescents and young adults in terms of a shift from conventional to postconventional moral thinking. However, we call our approach a "neo-Kohlbergian approach" (i.e., it is based on these starting points, but we have made some modifications in theory and method).

One major difference is our approach to assessment. Instead of Kohlberg's interview, which asks participants to solve dilemmas and explain their choices, the DIT1 uses a multiple-choice recognition task that asks participants to rate and rank a standard set of items. Some people are more accustomed to interview data and question whether data from multiple-choice tests are sufficiently nuanced to address the subtleties of morality research. Some researchers regard a multiple-choice test as a poor way to study morality, compared with the richness of explanation data from interviews. Therefore, the prior question concerning whether to update the DIT1 needs attention first. These challenges raise complex issues that are addressed in a recent book (Rest et al., 1999). Within the short span of an article, we can indicate only the general direction that we take.

The DIT Approach

A common assumption in the field of morality, and one with which we disagree, is that reliable information about the cognitive processes that underlie moral behavior is obtained only by interviewing people. The interview method asks a person to explain his or her choices. The moral judgment interview has been assumed to provide a clear window into the moral mind. In his scoring system (Colby et al., 1987), Kohlberg gave privileged status to interview data. At one point, Kohlberg (1976) referred to scoring interviews as "relatively error-free" and "theoretically the most valid method of scoring" (p. 47). According to this view, the psychologist's job is to create the conditions in which the participant is candid, ask relevant and clarifying questions, and then classify and report what the participant said. Then, in the psychologist's reports, the participant's theories about his or her own inner process are quoted to support the psychologist's theories of how the mind works.

However, consider some strange outcomes of interview material. When Kohlberg reported interviews, the participants talked like philosopher John Rawls (Kohlberg, Boyd,

& LeVine, 1990); when Gilligan reported interviews, the participants talked like gender feminists (Gilligan, 1982); and when Youniss and Yates (in press) reported interviews, the participants said that they don't reason or deliberate at all about their moral actions. This unreliability in explanation data exists because people do not have direct access to their cognitive operations. Perhaps people do not know how their minds work any more than they know how their immunization or digestive systems work. Perhaps asking a person to explain his or her moral judgments is likely to get back what they have understood current psychological theorists to be saying. Then, when psychologists selectively quote the participants' explanations that agree with their own views, such evidence is vulnerable to the charge of being circular. Thus, interview data need more than face validity.

Contrary to assuming the face validity of interviews, researchers in cognitive science and social cognition contend that self-reported explanations of one's own cognitive processes have severe limitations (e.g., Nisbett & Wilson, 1977; Uleman & Bargh, 1989). People can report on the products of cognition but cannot report so well on the mental operations they used to arrive at the product. We believe that people's minds work in ways they do not understand and in ways that they can't explain. We believe that one of the reasons that there is so little evidence for postconventional thinking in Kohlberg's studies (e.g., Snarey, 1985) is that interviewing people does not credit their tacit knowledge. There is now a greater regard for the importance of implicit processes and tacit knowledge in human decision making. Tacit knowledge is outside the awareness of the cognizer (e.g., Bargh, 1989; Holyoak, 1994) and beyond his or her ability to articulate verbally. For example, consider the inability of a 3-year-old to explain the grammatical rules used to encode and decode utterances in his or her native language. The lack of ability to state grammatical rules does not indicate what children know about language. Similarly, a lack of introspective access has been documented in a wide range of phenomena, including attribution studies (e.g., Lewicki, 1986), word recognition (Tulving, Schacter, & Stark, 1982), conceptual priming (Schacter, 1996), and expertise (Ericsson & Smith, 1991). This research calls into question the privileged place of interview data over recognition data (as in the DIT1). We believe that any data-gathering method needs to build a case for its validity and usefulness.

Note that the issue here is not whether Kohlberg distinguished normative ethics from meta-ethics. Rather, our point is that Kohlberg regarded explanation data from interviews as directly revealing the cognitive operations by which moral judgments are made. We are denying that people have access to the operations or inner processes by which they make moral decisions. We are denying that the royal road into the moral mind is through explanation data given in interviews. The upshot of all of this is extensive (see more detailed discussion in Rest et al., 1999). It not only means that multiple-choice data may have something of value to contribute to moral judgment research, but it also results in drawing the distinction between content and structure at a different place than Kohlberg did. All content is not purged from structure in assessment, the highest development in moral judgment is not defined in terms of a particular moral philosopher (i.e., John Rawls), and the concept of development is redefined so that development is not tied to the staircase metaphor.

We grant that the DIT started out in the 1970s as a "quick and dirty" method for assessing Kohlberg's stages. However, as time has passed and as data on the DIT1 has accumulated, different theories about human cognition have evolved (e.g., Taylor & Crocker, 1981). In keeping with these changes, we have reconceptualized our view of the DIT1 (see Rest et al., 1999, chapter 6). Now, we regard the DIT1 as a device for activating moral schemas (to the extent that a person has developed them) and for assessing them in terms of importance judgments. The DIT1 has dilemmas and standard items; the participant's task is to rate and rank the items in terms of their moral importance. As the participant encounters an item that both makes sense and also taps into the participant's preferred schema, that item is judged as highly important. Alternatively, when the participant encounters an item that either doesn't make sense or seems simplistic and unconvincing, he or she gives it a low rating and passes over it. The items of the DIT1 balance bottom-up, data-driven processing (stating just enough of a line of argument to activate a schema) with top-down, schema-driven processing (stating a line of argument in such a way that the participant has to fill in the meaning from schemas already in his or her head). In the DIT1, we are interested in knowing which schemas the participant brings to the task. We assume that those are the schemas that structure and guide the participant's thinking in decision making beyond the test.

Validity of the DIT1

Arguing that there are problems with interview data does not automatically argue for the validity of the DIT1. Rather, the DIT1 must make a case for validity on its own. Validity of the DIT1 has been assessed in terms of seven criteria. Rest, Thoma, and Edwards (1997) described the seven criteria for operationalizing construct validity. A recent book (Rest et al., 1999) cited over 400 published articles that more fully document the validity claims. The validity criteria briefly are as follows:

1. *Differentiation of various age and education groups.* Studies have shown that 30% to 50% of the variance of DIT scores is attributable to the level of education in heterogeneous samples.

2. *Longitudinal gains.* A 10-year longitudinal study showed significant gains of men and women and of college attenders and noncollege participants from diverse backgrounds. A review of a dozen studies of freshman to senior college students ($n = 755$) showed effect sizes of .80 (large gains). Of all the variables, DIT1 gains have been one of the most dramatic longitudinal gains in college (Pascarella & Terenzini, 1991).

3. DIT1 scores are significantly *related to cognitive capacity measures* of moral comprehension ($r = .60$s), recall and reconstruction of postconventional moral argu-

ments (Narvaez, 1998), to Kohlberg's moral judgment interview measure, and (to a lesser degree) to other cognitive developmental measures.

4. DIT1 scores are *sensitive to moral education interventions.* One review of over 50 intervention studies reported an effect size for dilemma discussion interventions to be .41 (moderate gains), whereas the effect size for comparison groups was only .09 (small gains).

5. DIT1 scores are significantly *linked to many prosocial behaviors and to desired professional decision making.* One review reported that 32 of 47 measures were statistically significant (see also Rest & Narvaez, 1994, for recent discussions of professional decision making).

6. DIT1 scores are significantly *linked to political attitudes and political choices.* In a review of several dozen correlates with political attitude, DIT1 scores typically correlated in the range, $r = .40$ to .60. When combined in multiple regression with measures of cultural ideology, the combination predicted up to two thirds of the variance (Rs in the .80s) of controversial public policy issues such as abortion, religion in the public school, women's roles, rights of the accused, rights of homosexuals, and free-speech issues. Such issues are among the most hotly debated issues of our time, and DIT1 scores are a major predictor to these real-life issues of macromorality.

7. *Reliability.* Cronbach's alpha is in the upper .70s/low .80s. Test–retest is about the same.

A specification of validity criteria tells us which studies to do to test a construct and what results should be found in those studies. Operational definitions enable us to examine the special usefulness of information from a measure. We want to know how the construct is different from other theoretically related constructs. Accordingly, DIT1 scores show *discriminant* validity from verbal ability and general intelligence and from conservative and liberal political attitudes. That is, the information in a DIT1 score predicts the seven validity criteria above and beyond that accounted for by verbal ability and general intelligence or political attitude (Thoma, Narvaez, Rest, & Derryberry, in press). Further, the DIT1 is equally valid for men and women (Rest et al., 1999). In sum, there is no other variable or construct that accounts as well for the combination of the seven validity findings than the construct of moral judgment. The persuasiveness of the validity data comes from the combination of criteria that many independent researchers have found, not just from one finding with one criterion.

Why a Revised DIT?

Because we wanted to maintain comparability in studies, DIT1 went unchanged while we went through a full cycle of studies. It took much longer to go through a full cycle than we originally anticipated; the DIT1 was frozen for over 25 years.

There are several issues about DIT1 that DIT2 seeks to address (and this moves us to the specific purposes of the present article):

1. Some of the dilemmas in DIT1 are dated, and some of the items needs new language (e.g., in DIT1, Kohlberg's well-known dilemma about "Heinz and the drug" is used, the Vietnam War is talked about in one dilemma as if it is a current event, and, in one of the items, the term *Orientals* was used to refer to Asian Americans). While updating dilemmas and items, we rewrote the instructions to clarify them, and we shortened the test from six stories to five stories when we found that one dilemma in DIT1 was not contributing as much to validity as were the other dilemmas (Rest, Narvaez, Mitchell, & Thoma, 1998b).

2. DIT2 takes advantage of a recently discovered way to calculate a developmental score (the N2 index; Rest, Thoma, Narvaez, et al., 1997). (Because issues of indexing are discussed at length in this recent publication, that discussion is not repeated here.)

3. There is the ever-present problem in group-administered, multiple-choice tests (that are also often anonymous) that participants might give bogus data. The challenge, therefore, is to develop methods for detecting bogus data so that we can purge the questionnaires that have bogus data. In DIT1, there are several checks for participant reliability; the usefulness of having some sort of check for participant reliability has been described (Rest, Thoma, & Edwards, 1997). Nevertheless, with DIT2, we reconsidered our particular method of checking for participant reliability, especially because such a large percentage (typically over 10%) of samples using DIT1 are discarded for questionable participant reliability. (Maybe in our zeal to detect bogus data, we threw out too many participants.)

To prepare the new dilemmas and items of DIT2, we first discussed various versions amongst ourselves. Then we asked members of an advanced graduate seminar on morality research at the University of Minnesota to take the reformulated DIT2 and to make comments. Then we discussed the dilemmas, items, and instructions again. Given that DIT1 has been unchanged for over 25 years and the fact that the Kohlberg group labored for decades over the scoring system of the moral judgment interview (Colby et al., 1987), changing the DIT might seem to be a big undertaking. However, the process was surprisingly straightforward and swift (and the results were positive). We conclude there is nothing sacred or special about the original Kohlberg dilemmas or the DIT1 dilemmas that cannot be reproduced in new materials. After freezing the DIT1 for years, we now encourage experimentation in new dilemmas and new formats. To encourage this experimentation, the new scoring guides and computer scoring from the Center for the Study of Ethical Development provide special aids to assist in the development of new dilemmas and new indexes (see Rest & Narvaez, 1998; Rest, Narvaez, Mitchell, & Thoma, 1998a).

DIT2 parallels DIT1 in construction:

1. Paragraph-length hypothetical dilemmas are used, each followed by 12 issues (or questions that someone deliberating on the dilemma might consider) representing different stages or schemas. The participant's task, a recognition task, is to rate and rank the items in terms of their importance.

2. The "fragment strategy" is used whereby each item is short and cryptic, presenting only enough verbiage to convey a line of thinking, not to present a full oration

defending one action choice or another (see Rest et al., 1999; Rest, Thoma, & Edwards, 1997).

3. Dilemmas and items on DIT2 closely parallel the moral issues and ideas presented in DIT1; however, the circumstances in the dilemmas and wording are changed, and the order of items is changed.

4. We presume that the underlying structure of moral judgment assessed by the DIT consists of three developmental schemas: personal interest, maintaining norms, and postconventional (Rest et al., 1999). See the Appendix for a sample story from DIT2.

Validating DIT2

How does one determine whether a new version of the DIT is working? We administered both DIT1 and DIT2 to the same participants, balancing the order of presentation. We included students at several age and education levels (from ninth-grade to graduate and professional school students). We wanted to pick criteria for this preliminary validation on which DIT1 was particularly strong, thinking that DIT2 would have to be at least as strong on these criteria. We used four criteria for initial validity:

1. *Discrimination of age and education groups.* This is our chief check on the presumption that our measure of moral judgment is measuring cognitive advance—a key assumption of any cognitive developmental measure.

2. *Prediction of opinions on controversial public policy issues.* As discussed in Rest et al. (1999), one of the most important payoffs of the moral judgment construct is its ability to illuminate how people think about the macromoral issues of society. The DIT predicts how people think about the morality of abortion, religion in public schools, and so on (matters dealing with the macro-issues of social justice, that is, how it is possible to organize cooperation on a society-wide basis, going beyond face-to-face relationships). The significant correlation between the DIT and various measures of political attitude has long been noted (see the review of over 30 correlations in Rest et al., 1999). A secondary goal of this study was to replicate a study by Narvaez, Getz, Thoma, and Rest (1999) by (a) using the specific measure of political attitude—the Attitudes Toward Human Rights Inventory (ATHRI; Getz, 1985); (b) testing whether DIT scores reduce to political identity or religious fundamentalism or to a common factor of liberalism or conservatism; and (c) testing whether or not the combination of DIT scores with cultural ideology (e.g., political identity and religious fundamentalism) more powerfully predicts controversial public policy issues than any one of these measures alone. Replicating the Narvaez et al. (1999) findings (both with DIT1 and DIT2 in a new study) is the first direct replication of these findings beyond the original study, on which we base our interpretation that moral judgment interacts with cultural ideology in parallel—not serially—in producing moral thinking about macromoral issues. More generally, we have taken the position that an important payoff of moral judgment research is to illuminate people's opinions about controversial public policy issues, and thus it is important to show that this interpretation is not based on only one study.

3. *High correlations between DIT1 and DIT2.* Of course this is important when comparing two tests purported to measure the same thing.

4. *Adequate internal reliability in DIT2.* This was the final criterion for determining the adequacy of DIT2.

We present our findings in four parts. Part 1 compares the performance of DIT2 (including the changes in dilemmas and items, in indexing, and in participant reliability checks) with DIT1, focusing on the four validity criteria mentioned previously. The central questions here are whether updating, clarifying, and shortening DIT2, and purging fewer participants for questionable reliability (practical improvements) can be done without sacrificing validity, and whether improvements in constructing a new index (N2) and new methods of detecting bogus data (new checks) are effective. In Part 2, we seek to isolate the effects of each of the three changes. What are the particular effects of changing the dilemma and item stimuli, the method of indexing, and the method of checking for participant reliability? In Part 3, we shift our focus to consider in some detail the problem of bogus data and methods for detecting unreliable participants. (New checks turns out to be the most unique methodological feature discussed in this article.) Finally, in Part 4, we further examine a replication with DIT2 of the Narvaez et al. (1999) study that concerns the discriminability of the DIT1 from political attitudes and examines the particular usefulness of the DIT2 in predicting opinions about public policy issues (seeking replication of the theoretical claim that moral judgment's most important payoff is the prediction of opinions about controversial public policy issues).

Method

Participants

The overall goal in constituting this sample was to have a mix of participants at various age and educational levels. We sought participants from four educational levels: students who were in the ninth grade, students who had recently graduated from high school and were enrolled for only a few weeks as freshmen in college, students who were college seniors, and students in graduate or professional school programs beyond the baccalaureate degree. These four levels of education have been used in studies of the DIT since 1974 (Rest, Cooper, Coder, Masanz, & Anderson, 1974). A total of 200 participants from these four age and educational levels turned in completions of all the major parts of the questionnaire package. Note that both the least advanced and the most advanced groups were from the upper Midwest, whereas the two middle groups were from the South. Thus, correlations with education could not be explained as regional differences.

Ninth-grade students. Two classrooms of ninth graders ($n = 47$) were asked to participate. The students attended a school that was located in a middle-class suburb of the Twin Cities metropolitan area. Testing took place over two class periods of a life skills class.

Senior high graduates, new freshmen. Students ($n = 35$) from a university in the southeastern United States were offered extra credit in several psychology classes for participation. Freshman students had recently graduated from high school and had been at the university for only a few weeks.

College seniors. Students ($n = 65$) from a university in the southeastern United States were offered extra credit in several

psychology classes. College seniors were students who were finishing their last year as undergraduates.

Graduate school and professional school students. Participants in this category (n = 53) consisted of 37 students in a dentistry program at a state university in the upper Midwest (at the end of their professional school program), 13 students at a private, moderately conservative seminary in the upper Midwest, and 3 students in a doctoral program in moral philosophy (we were unsuccessful in our attempts to recruit more moral philosophy students). Participants who took the tests on their own time were paid.

Instruments

The choice of instruments followed from the goals of the study, which are (a) to compare DIT1 with DIT2 and (b) to replicate the Narvaez et al. (1999) study.

Moral judgment: DIT1-P.[1] The DIT (Rest et al., 1999) is a paper-and-pencil test of moral judgment. DIT1 presents six dilemmas: (a) "Heinz and the drug" (whether Heinz ought to steal a drug for his wife who is dying of cancer, after Heinz has attempted to get the drug in other ways); (b) "escaped prisoner" (whether a neighbor ought to report an escaped prisoner who has led an exemplary life after escaping from prison); (c) "newspaper" (whether a principal of a high school ought to stop publication of a student newspaper that has stirred complaints from the community for its political ideas); (d) "doctor" (whether a doctor should give medicine that may kill a terminal patient who is in pain and who requests the medicine); (e) "webster" (whether a manager ought to hire a minority member who is disfavored by the store's clientele); and (f) "students" (whether students should protest the Vietnam War). Each dilemma is followed by a list of 12 considerations in resolving the dilemma, each of which represent different types of moral thinking. Items are rated and ranked for importance by the participant. For over 25 years, the most widely used index of the DIT1 has been the P score, representing the percentage of postconventional reasoning preferred by the respondent. Although the stages of moral thinking reflected on the DIT were inspired by Kohlberg's (1976) initial work, the DIT is not tied to a particular moral philosopher (as Kohlberg's is tied to Rawls, 1971). Kohlberg's stages are redefined in terms of three schemas (personal interests, maintaining norms, and postconventional).

DIT2–N2. The revised test consists of five dilemmas: (a) "famine" (A father contemplates stealing food for his starving family from the warehouse of a rich man hoarding food—comparable to the Heinz dilemma in DIT1); (b) "reporter" (A newspaper reporter must decide whether to report a damaging story about a political candidate—comparable to the prisoner dilemma in DIT1); (c) "school board" (A school board chair must decide whether to hold a contentious and dangerous open meeting—comparable to the newspaper dilemma in DIT1; (d) "cancer" (A doctor must decide whether to give an overdose of a painkiller to a frail patient—comparable to the doctor dilemma in DIT1; and (e) "demonstration" (College students demonstrate against U.S. foreign policy—comparable to the students dilemma in DIT1). The validity of DIT2 is unknown because this is the first study to use it. The N2 index takes into account preference for postconventional schemas and rejection of less sophisticated schemas, using both ranking and rating data. Its rationale is discussed in Rest, Thoma, and Edwards (1997).

Opinions about public policy issues. As in the Narvaez et al. (1999) study, the ATHRI, constructed by Getz (1985), asks participants to agree or disagree (on a 5-point scale) with statements about controversial public policy issues such as abortion, euthanasia, homosexual rights, due process rights of the accused, free speech, women's roles, and the role of religion in public schools. The ATHRI poses issues suggested by the American Constitution's Bill of Rights, similar to the large-scale studies of American attitudes about civil liberties by McClosky and Brill (1983). The ATHRI contains 40 items, 10 of which are platitudinous, "apple pie" statements of a general nature with which everyone tends to agree. Here are two examples of the platitudinous, noncontroversial items: "Freedom of speech should be a basic human right" and "Our nation should work toward liberty and justice for all." In contrast, 30 items are controversial, specific applications of human rights. Two examples are "Books should be banned if they are written by people who have been involved in un-American activities" and "Laws should be passed to regulate the activities of religious cults that have come here from Asia." During initial validation, a pro-rights group (from an organization that had a reputation for backing civil liberties) and a selective-about-rights group (from a group with a reputation for backing rights of certain groups selectively) were enrolled for a pilot study (n = 101) with 112 controversial items (Getz, 1985). Thirty of the items that showed the strongest divergence between groups were selected for the final version of the questionnaire, along with 10 items that expressed platitudes with which there was no disagreement (see Getz, 1985, for further details on the pilot study). Therefore, with the ATHRI, we have a total of 40 human rights issues that are related to civil libertarian issues.

In the study by Narvaez et al. (1999), scores ranged from 40 to 200. These high scores represent advocacy of civil liberties. Although the items of the ATHRI represent many different issues and contexts, they strongly cohere (Cronbach's alpha was .93). Narvaez et al. (1999) reported significant bivariate correlations of DIT1 with ATHRI (rs in the .60s). Also, when measures of political identity and religious fundamentalism were combined in multiple regression with the DIT to predict ATHRI, the R was in the range of .7 to .8, accounting for as much as two thirds of the variance. Further, each of the independent variables had unique predictability (as well as shared variance). Thus, each independent variable was not reduced to a single common factor of liberalism or conservatism. The present study was intended to replicate those findings using a different sample, with both DIT1 and DIT2.

Religious ideology. To measure religious fundamentalism, we chose Brown and Lowe's (1951) Inventory of Religious Belief, following Getz (1985) and Narvaez et al. (1999). It is a 15-item measure that uses a 5-point, Likert-type scale. Its items differentiate between those who believe and those who reject the literalness of Christian tenets. It includes items such as "I believe the Bible is the inspired Word of God" (a positively keyed item); "The Bible is full of errors, misconceptions, and contradictions" (a negatively keyed item); "I believe Jesus was born of a virgin"; and "I believe in the personal, visible return of Christ to earth." Scores on the Brown–Lowe inventory range from 15 to 75. High scores indicate strong literal Christian belief. Criterion group validity is good between more and less fundamentalistic church groups (Brown & Lowe, 1951; Getz, 1984; Narvaez et al., 1999). Test–retest reliability has been reported in the upper .70s. Spearman–Brown reliability has been found in the upper .80s (Brown & Lowe, 1951). In Narvaez et al. (1999), Cronbach's alpha was .95 for the entire

[1] Operationalized variables used in statistical analysis are printed as an abbreviated name in capital letters (e.g., DIT1-P, FUNDA). Theoretical constructs are printed in the usual manner (e.g., moral judgment, religious fundamentalism). In the case of DIT variables, the version is designated by DIT1 or DIT2, and the index used is designated after the hyphen (e.g., DIT1-P, the original DIT using the P index; or DIT2-N2, the new DIT using the N2 index).

Table 1
Participants Groups and Demographics

Group	Number	Average age (*SD*)	Percent women
Ninth grade	47	14.64 (0.53)	34
High school graduates/college freshmen	35	18.51 (2.03)	77
College seniors	65	21.55 (3.11)	77
Graduate/professional school	53	29.06 (5.90)	45
Total	200	21.4 (6.39)	58.5

group of 158 participants. This scale taps religious fundamentalism and is labeled FUNDA.

Political identity: Liberalism and conservatism. Participants were asked to identify their political identity on a 5-point political conservatism scale, ranging from 1 (*liberal*) to 5 (*conservative*). This method of measuring liberalism and conservatism replicates the Narvaez et al. (1999) study and is the variable of contention in the challenge to the DIT1 by Emler, Resnick, and Malone (1983). This variable will be referred to as POLCON (political conservatism), with high scores being conservative.

Demographics. Age of participants was given in years. Participants were also asked to state their gender, but because there were no significant differences on any of the DIT scores for gender scores for both males and females were collapsed for analysis. Education was measured in terms of the four levels of education (1 = ninth grade, 2 = college freshman, 3 = college senior, 4 = graduate or professional school student). Participants were asked whether they were Christians. Participants were also asked whether they were citizens of the United States (virtually all, 98.3%) and whether English was their first language (virtually all, 97%). Some participant demographics are shown in Table 1.

Procedure

The order of materials was randomly varied (for all but the dentistry students), with DIT1 coming first for half of the participants and DIT2 coming first for the other half. There were no significant differences in terms of order for any of the major variables (P and N2 indexes on DIT1, P and N2 on DIT2 or on ATHRI, FUNDA, and POLCON). Because the 37 dentistry students had already taken the DIT1 as part of their regular curriculum requirements, we sought volunteers to take the remaining package of questionnaires, and the order was not varied.

For the high school participants, time in two class sessions was used to take the questionnaires; for the remaining participants, the questionnaire package was handed out and the participants filled out the questionnaires on their own time.

All Minnesota participants (and parents of the ninth graders) signed consent forms in accordance with the procedures of the University of Minnesota Human Participants Committee. Participants from the southeastern university were recruited in compliance with that institution's human participant requirements.

Results and Discussion

In Part 1, DIT1-P is compared with DIT2-N2. How does the new revision of the DIT stack up against the traditional DIT, which has been used for over 25 years and reported in hundreds of studies? The key question is whether, after decades of research, we have developed a better instrument. Then in Part 2, we examine the particular effects of each of the three changes in DIT2: (a) using the original wording of dilemmas and items versus the revised dilemmas and items, (b) using the P index versus using the new N2 index, and (c) using the standard participant reliability checks versus using new checks.

Part 1

Participant reliability. The DIT contains checks on the reliability of a participant's responses. DIT1 uses a different method for detecting participant unreliability than the DIT2 (discussed in detail in Part 3). From the total sample of 200 participants, 154 survived the reliability checks of the standard procedure for DIT1 (77%), whereas 192 survived the new reliability checks of DIT2 (96%). Given that in this study the same participants took both DIT1 and DIT2, we conclude that DIT2 purges fewer participants for suspected unreliability than does DIT1. The difference in proportion of participants purged between the new procedure and the standard procedure is significant ($z = 5.56, p < .0001$).

Criterion 1. We expect a developmental measure of moral judgment to increase as age and education increases. Table 2 presents the means and standard deviations of DIT1-P and DIT2-N2 for each of the four educational levels. An analysis of variance (ANOVA) with DIT1-P grouped by four levels of education produces $F(3, 153) = 41.1$, $p < .0001$; an ANOVA with DIT2-N2 produces $F(3, 191) = 58.9$; $p < .0001$. Table 3 presents age and educational trend data in terms of correlations of the moral judgment indexes with educational level (four levels) and with chronological age (14–53). Although there might be doubts about the strict linearity of education level (and therefore the use of level of education as a linear variable in correlations), we assume that deviations of the educational-level variable from strict linearity affects both DIT1 and DIT2 equally, thus not biasing the comparison between DIT measures. The correlational analysis shows stronger educational trends with DIT2-N2 than with DIT1-P, although this amount of difference may not make much practical difference. In sum, the practical advantages of DIT2 (i.e., being shorter, more up-to-date, and purging fewer participants) are not at the

Table 2
Means and Standard Deviations of DIT1-P and DIT2-N2 by Four Education Levels

	DIT1-P (n = 154)		DIT2-N2 (n = 192)	
Education level	*M*	*SD*	*M*	*SD*
1. Ninth grade	23.0	10.0	20.5	9.7
2. College freshmen	28.7	11.5	30.6	14.4
3. College seniors	33.7	14.1	40.4	13.6
4. Graduate/professional school	53.9	13.1	53.3	11.5

Note. In DIT1-N2, for comparison purposes, the N2 index is adjusted so that the mean (37.85) and standard deviation (17.19) are equal to those of the P index. DIT1 = Defining Issues Test (original version); DIT2 = Defining Issues Test, Version 2; P = P index; N2 = N2 index.

Table 3
Correlations of DIT With Education and Age

Measure	Education level (1–4)	Chronological age
DIT1-P index	.62	.52
DIT2-N2 index	.69	.56

Note. All correlations of DIT with age and education level are significant, $p < .0001$. The correlation of DIT2-N2 with education level is significantly higher, $t(151) = 6.72$, $p < .001$, than the correlation of DIT1-P with educational level. The correlation of DIT1-P with age is not significantly different, $t(151) = 1.67$, *ns*, from the correlation of DIT1-N2 with age. (Calculation of differences between correlations follows Howell, 1987, pp. 244ff, in which the correlations are first transformed to Fisher's *r*.) DIT1 = Defining Issues Test (original version); DIT2 = Defining Issues Test, Version 2; P = P index; N2 = N2 index.

cost of poorer validity on Criterion 1. In fact, the opposite is true.

Criterion 2. We expect a measure of moral judgment to be related to views on public policy issues such as abortion, free speech, rights of homosexuals, religion in public schools, women's roles, and so on.

In Table 4, we show both the old and new DIT correlated with ATHRI and also the partial correlations with ATHRI controlled for FUNDA and POLCON. We show partial correlations because previous studies (Rest, 1986) have shown that both religious fundamentalism and political conservatism and liberalism were significantly correlated with the DIT. Therefore, the partial correlation attempts to control the shared variance with political or religious conservatism of the DIT with ATHRI, estimating the relation of moral judgment to public policy issues after controlling for religious and political conservatism. Again, despite the practical advantages of DIT2-N2 over DIT1-P, the new version does not suffer any weaker trends on Criterion 2. In fact, in the partial correlation with ATHRI, DIT2-N2 has a significant advantage over DIT1-P.

Criterion 3. We expect a measure of moral judgment to have adequate reliability as measured by Cronbach's alpha.

Table 4
Correlations and Partial Correlations of Moral Judgment With ATHRI

Measure	ATHRI	ATHRI (controlling for FUNDA and POLCON)
DIT1-P	.48	.40
DIT2-N2	.50	.51

Note. All correlations of the DIT with ATHRI are significant, $p < .001$. The correlation of DIT1-P with ATHRI is not significantly lower, $t(151) = .99$, *ns*, from the correlation of DIT1-N2 with ATHRI. The correlation of DIT1-P with ATHRI, partialing out for FUNDA and POLCON, is significantly lower, $t(149) = 4.43$, $p < .001$, than the corresponding partial correlation of DIT2-N2. ATHRI = Attitudes Toward Human Rights Inventory (Getz, 1985); FUNDA = religious fundamentalism; POLCON = political identity as conservative; DIT1 = Defining Issues Test (original version); DIT2 = Defining Issues Test, Version 2; P = P index; N2 = N2 index.

Because we use ranking data in the P index and as part of the N2 index, we cannot use the individual items as the unit of internal consistency. Ranks are ipsative; that is, if one item is ranked in first place, then no other item of a story can be in first place. Therefore, the unit of internal reliability is on the story level, not the item level. Cronbach's alpha for DIT1-P over the six stories ($n = 154$) is .76. For the DIT2-N2, it is .81 ($n = 192$). Although these levels of Cronbach's alpha are not outstandingly high, we regard them as adequate.

It is interesting to note that Cronbach's alpha for DIT1's 6 stories plus DIT2's 5 stories (for a total of 11 units) is .90. We might speculate that this finding (i.e., 5 or 6 stories have modest reliability, but 11 stories have high Cronbach's alpha) indicates that the five or six stories of DIT1 and DIT2 each tap some different subdomains within morality. Although the DIT1-P and DIT2-N2 cohere enough, there is nevertheless some diversity in what each story taps. When we add the 6-story DIT1 to the 5-story DIT2, the 11 stories show higher internal consistency because the 11 stories have more overlap and are more redundant than the smaller samples of the 5 or 6 stories. Paradoxically, however, a score based on the 11 stories contains essentially the same information (although somewhat redundantly) as the score from 5 stories (with less redundancy). This can be seen by comparing the correlations of the validity criteria from the 5-story DIT2-N2 with the 11-story DIT1 + DIT2: For the 5-story DIT2-N2, the correlation with education is .69, whereas the correlation with the 11-story DIT is .73. The correlation of the 6-story DIT2-N2 with ATHRI is .50, whereas the correlation of the 11-story DIT1 + DIT2 is .52. By using all 11 stories (virtually doubling the test), the gain in Cronbach's alpha is 8 points, whereas the gain in the correlations with validity criteria is only 2 to 4 points. (Hence, we conclude that on the basis of 5 stories, DIT2-N2 contains virtually the same information as a moral judgment variable that is based on 11 stories with high Cronbach's alpha.)

Criterion 4. We expect DIT1 to be significantly correlated with DIT2. This criterion is different from the previous three criteria in that it does not contrast DIT1 with DIT2, but, rather, examines their overlap. The correlation of DIT1-P with DIT2-N2 is .71 (using the standard participant reliability checks; $n = 154$). The correlation of DIT1-N2 with DIT2-N2 is .79 (using the N2 index and new checks; $n = 178$).

With Guilford's (1954, p. 400) correction for attenuation resulting from the less-than-perfect reliability of two measures, the upward bound estimate for the correlation between the two "true" scores is .95 to .99 (depending on the sample used for reliability estimates and the method of indexing). Hence, the DIT1 and DIT2 are correlated with each other about as much as their reliabilities allow. DIT1 is correlated with DIT2 about as much as previous studies have reported for the test–retest of DIT1 with itself (Rest, 1979, p. 239).

In sum, DIT2-N2 is shorter, more streamlined, more updated, and purges fewer participants than DIT1-P, and (with N2 and new checks) it has somewhat better validity characteristics. According to this study, if either measure has

the validity advantage, it seems to lie with DIT2 in addition to its practical advantages.

Part 2

What effects are unique to the new dilemmas and items and what effects are a result of the new analyses (N2 and new checks)? What if the new analyses are computed on the old DIT (i.e., the data from the 6-story DIT1)? Would there be any advantage in doing so (without using the new dilemmas and items)?

In Table 5, the top row repeats the correlations of DIT2-N2 with the validity criteria already given in Tables 2 and 3 and in the discussion of Cronbach's alpha in *Criterion 3;* the bottom row repeats the correlations of DIT1-P with the validity criteria (also given previously). Rows 1 and 4 are provided for easy comparison with rows 2 and 3. The second row (the most important row in Table 5) shows how the validity criteria are affected by using the old DIT (Heinz and the drug, etc.) with the new index (N2) and the new participant reliability checks. (In other words, row 2 uses the old DIT, including Heinz and the drug, but adopts the new data analyses of N2 and new checks.) The special interest in row 2 is whether there seems to be any advantage to reanalyzing DIT1 with N2 and new checks. The third row shows how the correlations are affected by using the new DIT with the old P index and the old standard reliability checks.

First, note the sample sizes. The new participant reliability checks allow more participants in the sample to be cleared for analysis (96% for DIT2; 92% for DIT1) than do the standard reliability checks (77% for DIT-P). The difference between 92% and 77% is statistically significant ($z = 3.98$, $p < .001$, $n = 200$), and the difference between 96% and 92% is statistically significant ($z = 1.86$, $p = .05$, $n = 200$). In other words, the new analyses (N2, new checks) retain significantly more participants on both DIT1 and DIT2 than do the old standard analyses, and with new checks, DIT2 retains slightly more participants than does DIT1. Although new checks retains more participants than does standard checks on both DITs, DIT1 lost nine more participants than did DIT2 using new checks.

Second, note that using the new analyses (N2 and new checks) makes more of a difference in the validity criteria than using new dilemmas (DIT2). In other words, the old DIT (6-story DIT1)—for all its datedness and awkward wording—seems to produce trends as strong as the new DIT (5-story DIT2) with updated wording when the new analyses (N2 and new checks) are used. The particular advantages of DIT2 seem mostly to be that it is shorter and retains slightly more participants (nine more than DIT1), not that the changes in dilemmas or wording produce stronger validity trends. Perhaps the datedness and awkward wording of DIT1 put off some participants and undermined motivation to perform the task, but in the current study, this seemed to affect only 5% of the participants. When most participants perform the task of DIT1, the validity trends are as strong as the updated, shorter version. In both cases, however, the new analyses with N2 and new checks are preferable to the analyses used for over 25 years for DIT1.

The third row in Table 6 shows that it is not a good idea to use DIT2 without the N2 index and new checks. From the perspective of this study, the only disadvantage of N2 and new checks is that they are too labor intensive for hand scoring (the original DIT1 could be hand scored). It takes several hours of hand computation per participant to perform the routines of N2 and new checks. Only a computer should be put through the amount of calculation necessary to produce N2 and new checks.

One might wonder whether the DIT's relation to ATHRI is "piggybacking" on a third variable, education. After all, other research (e.g., McClosky & Brill, 1983) has shown correlations between public policy issues and education. Therefore, partialling out education, the partial correlation of DIT1 with ATHRI is .30 ($n = 180$, $p < .001$), and the partial correlation of DIT2 with ATHRI is .28 ($n = 195$, $p < .001$). Again, partialling out education, the partial correlation of DIT1 with DIT2 is .62 ($n = 178$, $p < .001$). Therefore, there is no indication that education can account for the predictability of the DIT.

Table 5
Correlations of DIT Measures With the Validity Criteria With and Without New Index and New Participant Reliability Checks

Measure	n[a]	ED[b]	ATHRI[c]	ATHRI/partial[d]	Cronbach's α
With new index and new participant reliability checks					
DIT2-N2	192	.69	.50	.51	.81
DIT1-N2	183	.68	.54	.52	.81
With old P index and standard participant reliability checks					
DIT2-P	154	.62	.55	.42	.74
DIT1-P	154	.62	.48	.40	.76

Note. ED = educational level; ATHRI = Attitudes Toward Human Rights Inventory (Getz, 1985); DIT1 = Defining Issues Test (original version); DIT2 = Defining Issues Test, Version 2.
[a]Sample retained after participant reliability checks. [b]Correlation with educational level (4 levels). [c]Bivariate correlation with ATHRI. [d]Partial correlation with ATHRI, controlling for religious fundamentalism (FUNDA) and political identity as conservative (POLCON).

Table 6
Multiple Regressions of Moral Judgment, Political Identity, and Religious Fundamentalism (Independent Variables) Predicting Controversial Public Policy Issues (Dependent Variable)

Variable	*B*	β	*t*
Equation 1, predicting to ATHRI including DIT1-P, $n = 154$, multiple $R = .56$, $df = 151$			
DIT1-P	0.34	.38	5.3***
POLCON	−3.78	−.23	−3.2**
FUNDA	−0.18	−.15	−2.0*
Equation 2, predicting to ATHRI including DIT2-N2, $n = 194$, multiple $R = .58$, $df = 191$			
DIT2-N2	0.28	.48	8.0***
POLCON	−3.85	−.23	−3.7**
FUNDA	−0.17	−.14	−2.2*

Note. POLCON and FUNDA are negatively related because high scores are more conservative, in contrast to DIT and ATHRI scores, which run in the opposite direction. ATHRI = Attitudes Towards Human Rights Inventory. DIT1-P = Defining Issues Test (original version), using P index; POLCON = political identity as conservative; FUNDA = religious fundamentalism; DIT2-N2 = Defining Issues Test, Version 2, using N2 index.
*$p < .05$. **$p < .01$. ***$p < .001$.

For completeness of analysis, additional tables of the validity criteria were also computed to separate the effect of indexing from methods of detecting participant reliability (i.e., using the P index with new checks, and using N2 with standard checks). The results were generally intermediate between rows 1 and 4. So nothing of special interest was found here. The general conclusion is that, for the strongest validity trends, the researcher might use either DIT1 or DIT2, but should use both new analyses together. The practical advantages of DIT2 (i.e., it is somewhat shorter, less dated, and likely to retain slightly more participants) is what recommends it over DIT1. We were expecting that the new dilemmas and wording of DIT2 would make a contribution to greater validity (in addition to using N2 and new checks), but we were surprised that DIT1 seems to work about as well (when used in conjunction with N2 and new checks).

Part 3

Recall that DIT2 involves three changes from DIT1: (a) changes in dilemmas and items (discussed earlier); (b) changes in indexing (discussed in detail elsewhere; Rest, Thoma, Narvaez, et al., 1997); and (c) changes in participant reliability checks, which is addressed in this section.

The problems of bogus data. One inevitable problem with a group-administered, multiple-choice test is that participants might put check marks down on the questionnaire without reading the items or following instructions, or they might proceed with a test-taking set that is alien to the instructions. How do researchers determine whether the participants' responses reflect moral thinking (as the moral judgment construct purports) or are bogus? There are four problem responses that give bogus data:

1. *Random responding.* The participant may fill in the bubbles on an answer sheet, but the marks may not have anything to do with his or her moral cognition. For instance, we have seen answer sheets on which participants filled in the answer bubbles to form Christmas trees and other geometric designs. We doubt that such responses accurately measure the construct of moral judgment.

2. *Missing data.* The participant may not be sufficiently motivated to take the test and may leave out large sections of answers, or just quit.

3. *Alien test-taking sets.* The participant may choose items not on the basis of their meaning, but on the basis of complex syntax, special wording, or the seemingly lofty sound of the words. In this case, the scores do not reflect the moral judgment construct but instead reflect a preference for complex style or verbiage.

4. *Nondiscrimination of items.* The participant may put down the same response for all items, failing to discriminate among the items (e.g., putting down *3*s for all ratings and ranks). Rest, Narvaez, Mitchell, and Thoma (1998a) showed that for a very large sample ($n > 58{,}000$), participants show considerable variation in rating and ranking DIT items; therefore, some variation is expected.

If a participant is suspected of any one of these four response problems, we know of no way to salvage or correct the protocol. Instead, the entire protocol is discarded from analysis. In general, previous research has shown that purging the protocols of participants who manifest any of these four problems results in clearer data trends (Rest, Thoma, & Edwards, 1997), presumably because error variance has been minimized.

Standard checks. In the standard checks procedure used with DIT1, four checks are used to detect the likelihood of each of the four problems.

1. *The problem of random responding.* As a guard against random checking, a participant's ratings are checked for consistency with the participant's rankings. For example, if a participant chose Item 10 as the top rank (most important item), then no other item should be rated higher in importance than Item 10. Further, with this example, if a participant chose Item 8 as second most important rank, then only Item 10 should be rated higher. Our general approach is to count each violation of a pattern of rank–rate consistency as an inconsistency. Thus, with regard to the first problem (random responding), rate–rank inconsistencies are assumed to indicate random checking. Theoretically, the perfectly consistent participant will have no rank–rate inconsistencies. In reality, however, we can expect some inconsistency, even among serious, well-motivated participants. Participants sometimes change their minds after being exposed to a variety of issues. So the question becomes, how much inconsistency should researchers tolerate as the innocent shifting of item evaluations, and how much inconsistency is too much, reflecting random responding? Where do we draw the line? In the standard procedure, participants who have more than eight inconsistencies on a dilemma (counting only

the top two ranks) are considered to have too much inconsistency as are participants who have inconsistencies on more than two dilemmas. Participants exceeding these cutoff points are eliminated from the sample. It turns out, over our 25-year experience, that this rate–rank consistency check (more than the other three participant reliability checks, described later) accounts for the bulk of purged participants.

2. *The problem of missing data.* Occasional missing data are tolerated in standard checks. For example, if someone omits an occasional rating or ranking, we do not purge the entire protocol. Instead, we readjust scores to make up for the missing data, in effect calculating readjusted scores to reflect the response patterns in the rest of the protocol and adjusting scores so that every participant's data are on the same scale. However, too much missing data may reflect a general lack of motivation to take the task. In this case, we cannot have confidence in any responses. Again, the question is how much is tolerable and how much is grounds for purging the entire protocol? In the standard checks procedure, if a participant leaves out two whole stories (for instance, a participant is asked to complete six stories but only completes four), then that is regarded as too much. The problem in interpreting such a protocol is not that we could not readjust a score based on four stories to be on the same scale as six stories; rather, the problem is that we are suspicious of the motivation of the participant to do the work of the DIT in the four stories (it is possible that even in the four stories, reliable data were not given).

3. *The problem of alien test-taking sets.* Participants who choose items for their pretentiousness or lofty sound are not following instructions to choose items based on their meanings. As a check on this alien test-taking set, we have distributed five meaningless items (M-items) throughout the DIT that may be attractive for their complex syntax or "high sounding" verbiage, but do not mean anything. If a participant ranks too many of these M-items too highly, we assume an alien test-taking set and purge the whole protocol. In the standard procedure, a score of 8 or more (weighting ranks by 4 for top rank, by 3 for second rank, etc.) on the M-items invalidates the protocol.

4. *The problem of nondiscrimination.* Participants who do not discriminate answers (e.g., those who check *3*s for all items) are not complying with our instructions to make discriminations. Because nondiscriminating participants will not be picked up in the rate–rank check, a special check has been devised for nondiscrimination. In the standard procedure, no more than one story can have more than eight items rated the same.

New checks. The new checks procedure recognizes the same four problems in participant reliability, but deals with them in ways different from the standard checks procedure. To investigate the consequences of different methods and cutoff points, we concocted a set of protocols that deliberately epitomized one or more of the violations we sought to detect. Some of the deliberately bogus data were based on a random number table (to simulate random responding). Other bogus protocols were based on filling in the answer bubbles to form graphic designs (e.g., the Christmas tree design). In general, the objective was to have protocols that we knew were bogus data and to see whether our reliability checks would pick up all these bad protocols, but would pass through a high percentage of actual data. We also wanted to see if the validity trend was still robust with new cutoff scores. We were especially interested in comparing data trends of the new checks with the old standard checks.

1. *The problem of random responding.* Because in the standard checks, the largest numbers of participants are purged for unreliability based on the rate–rank consistency check, we paid the most attention to this procedure. To detect participants who are randomly checking, this is our new procedure: We look at a participant's ranks, weight the top rank as 4, the second most important as 3, the third as 2, and the fourth as 1 (same weights as in deriving the P score). Then we look at the item's rating. If there is an item different from the one in the first rank that is rated more highly than the item in the top rank, then that is one occurrence of inconsistency and is multiplied by 4. All other inconsistencies with the top rank are also multiplied by 4. Then we look at the item ranked as second most important. There should not be any item rated more highly than the second-ranked item except the item ranked in the top rank. The occurrences of exceptions to this expectation are counted and weighted by 3, and so on for the third- and fourth-ranked items (the violations are counted and weighted by 2, for third rank, or by 1, for fourth rank). The weighted inconsistencies for each story and across stories are summed. The summed weighted rank–rate inconsistencies across five stories can range from 0 to 600. Through trial and error, we arrived at cutoff points. We wanted a stringent enough threshold point to prevent any of the deliberately bogus data from getting through, but not so low a threshold to make the validity trends suffer. Thus, we arrived at the cutoff point of how much is too much by empirical trial and error. It turns out that if the sum of rate–rank inconsistencies is more than 200, then that is too much, and the protocol is invalidated (purged from the sample). If the sum is under the 200 mark, it is regarded as innocent confusion, and we tolerate that much inconsistency by not purging the entire protocol.

2. *The problem of missing data.* Occasional missing data are tolerated by DIT2. Using the trial-and-error procedure described earlier, we arrived at cutoff values. If the participant leaves out more than three ratings on any of two stories, the protocol is invalidated. If the participant leaves out more than six ranks, the protocol is also invalidated.

3. *The problem of alien test-taking sets.* Participants who pick items for style rather than for meaning are not following our instructions. In the new checks procedure, we also use M-items to detect this problem. The protocols of participants whose weighted ranks on the M-items total more than 10 are invalidated (more lax than the cutoff of 8 on standard checks).

4. *The problem of nondiscrimination.* In new checks, participants who rate 11 items the same on a story are considered as not discriminating; if the participant fails to discriminate on two stories or more, the protocol is invalidated. Nondiscrimination by rates or ranks is grounds for purging the protocol.

As mentioned earlier, the new checks purged 8 participants from the sample of 200 (or 4%), whereas the standard checks purged 46 participants (or 23%) from the sample. In general, the new checks are less stringent that the standard checks. Paradoxically, the data in this study suggest that the less stringent method (new checks) produces stronger trends than does the more stringent method (standard checks). How can this be? One might expect the opposite—that making sure of participant reliability (the more stringent method) would produce stronger validity trends than a more lax method for checking for participant reliability. The key to this paradox lies in the fact that standard checks purges proportionately more of the youngest group of ninth graders (58% of the ninth graders were purged by the standard checks) than for the oldest group (only 8% were purged in the graduate and professional school subsample). In contrast, new checks purged only 11% of the ninth graders (and 1 participant from the graduate school subsample). The difference between 58% and 11% is significant, $z = 4.80$, $p < .0001$, $n = 47$. One might speculate that with standard checks, the disproportionate purging of the youngest participants from the sample in effect changes the distribution of scores in the total sample, making the sample more homogeneous, attenuating the spectrum of scores, and resulting in slightly weaker correlations and validity trends for standard checks. In other words, the new checks have stronger validity trends because they retain more of the lower scores from the youngest participants and thus retain a wider range of scores (by which the correlations increase).

Because the cutoff values for the reliability checks are empirically derived, it remains to be seen whether they are optimum for other samples. The experiences of other researchers is the most important consideration here. To facilitate experimentation with different cutoff values for the checks, the scoring service of the Center for the Study of Ethical Development provides a set of variables that can be manipulated for each sample (Rest & Narvaez, 1998; Rest, Narvaez, Mitchell, & Thoma, 1998a).

Part 4

Recall that Criterion 2 of validity in this study deals with the correlation of the DIT with ATHRI. The correlation of moral judgment with political attitudes has been noted for some time (typically *r*s in the .4 to .6 range; see Rest et al., 1999, for a review of several dozen correlations over 25 years). Emler et al. (1983) interpreted this pattern of correlations, contending that the DIT is really liberalism–conservatism masquerading as developmental capacity. They stated that

> Moral reasoning and political attitude are by and large one and the same thing. . . . We believe that individual differences in moral reasoning among adults—and in particular those corresponding to the conventional-principled distinction—are interpretable as variations on a dimension of political-moral ideology and not as variations on a cognitive-developmental dimension. (pp. 1073–1075)

In contrast, our view (Narvaez et al., 1999) is that moral judgment, political identity (identifying oneself as a liberal or conservative), and religious fundamentalism are related but also distinct constructs. The variables carry unique information, and they cannot all be reduced to a common factor of liberalism–conservatism. (See Thoma et al., in press, and Rest et al., 1999, for discussion of the Emler et al., 1983, studies.) In support of the uniqueness of each construct or variable (moral judgment, religious fundamentalism, and political identity as liberal or conservative), Narvaez et al. reported a multiple regression having the ATHRI as the dependent variable and having the DIT, FUNDA, and POLCON as independent variables. Multiple regression analysis permits estimation of the unique contribution of each independent variable by examining the standardized beta weights. Narvaez et al. (Study 1) examined two church congregations and found that the beta weights for each of the three independent variables were each significant in their own right, indicating that each contributes distinct information to ATHRI. This finding for church samples was replicated in a student sample (Narvaez et al., 1999, Study 2). Now we wish to determine whether the findings replicate with both DIT1 and DIT2. Because we place so much importance on the DIT's unique contribution to understanding opinions about controversial public policy issues (the macrolevels of morality), we wanted to have more than just the Narvaez et al. studies to confirm our interpretation.

Because we wanted to replicate the Narvaez et al. (1999) study, we used the Brown and Lowe (1951) instrument as the measure of fundamentalism. However, there is a problem in that the Brown and Lowe instrument is a measure of Christian fundamentalism, and the sample from this study could have included Orthodox Jews, Orthodox Muslims, or others who would have a low score on Christian fundamentalism but nevertheless be very orthodox in a non-Christian way. Checking the total sample, it turned out that the overwhelming proportion (90%) indicated they considered themselves to be Christian. Only 20 participants indicated that they were non-Christian. However, leaving these participants out of the analysis made little difference in the relation of FUNDA to DIT, ATHRI, or POLCON. Correlations of FUNDA with DIT2-N2, ATHRI, and POLCON, *including* the non-Christians, were −.10, −.25, and .28, respectively; *excluding* the non-Christians, the correlations were −.13, −.26, and .21, respectively. Because including or excluding the non-Christians made little difference where FUNDA was concerned (Criterion 2), we left the non-Christians in the sample for the sake of maximizing the sample size on the other three criteria.

The multiple regressions in Table 6 on this sample replicate with DIT1-P and with DIT2-N2 in the Narvaez et al. (1999) studies: (a) Both studies used the same dependent variables, ATHRI (controversial public policy issues), and independent variables (FUNDA, POLCON, and DIT); (b) each independent variable (DIT, POLCON, and FUNDA) has significantly unique predictability to ATHRI; (c) moral judgment has higher standardized beta weights than does POLCON or FUNDA; (d) when all three independent variables are combined, the combination predicts powerfully

Table 7
Predictability to ATHRI From Multiple Regression Beta Weights From Present Study With Multiple Regression Weights From Narvaez et al. (1999), Study 1

Measure	ORTHO[a] (Study 1)	Multiple R[b] (present study with DIT1-P)
ATHRI (n = 154)	−.54	.56
	ORTHO[c] (Study 1)	Multiple R[d] (present study with DIT2-N2)
ATHRI (n = 192)	−.55	.58

Note. DIT, POLCON, and FUNDA are components that go into ORTHO and multiple R (from Table 5). All correlations are significant, $p < .001$. DIT1-P = Defining Issues Test (original version), using P index; DIT2-N2 = Defining Issues Test, Version 2, using N2 index; ATHRI = Attitudes Toward Human Rights Inventory (Getz, 1985).
[a]Orthodoxy combination variable formed by combining DIT1-P, POLCON, and FUNDA according to weights of multiple regression in Study 1 of Narvaez et al. (1999). [b]Multiple regression from Table 5, Equation 1. [c]Same combination of variables based on weights of Study 1, but with DIT2-N2 for moral judgment component. [d]Multiple regression from Table 5, Equation 2.

to ATHRI (with DIT1-P, R = .56; with DIT1-N2, multiple R = .58; with the 11-story DIT1 + 2 − N2, R = .63).

A stronger test of Narvaez et al. (1999) is to use the same beta weights (nonstandardized) as those in Study 1 for combining DIT, POLCON, and FUNDA. Using beta weights from the original multiple regression (in Narvaez et al., 1999, Study 1) produces a variable called ORTHO (representing the construct, orthodoxy–progressivism, as discussed by Hunter, 1991). ORTHO provides a stronger replication of the Narvaez et al. (1999) study than a new multiple regression on this new sample because in combining the three variables from the beta weights of the original sample, we are not capitalizing on sample-specific chance factors (as do the multiple regressions in Table 6). ORTHO is, in effect, a transfer of the original relations of the independent variables from Narvaez et al. (1999, Study 1) to the present study in predicting ATHRI. Table 7 shows that the correlation of ATHRI with ORTHO is only two or three points weaker than the Rs run in Table 6 on the specific new samples of the present study. In other words, the beta weights derived from the multiple regression for ORTHO from Study 1 in Narvaez et al. (1999) generalizes well to the present study.

One might be concerned with the problem of multicollinearity on the multiple regression results. As Howell (1982, pp. 500ff) noted, a problem can exist in interpretation of multiple regression results when the independent variables are intercorrelated; the problem is that the beta weights are unstable from sample to sample.

In the present sample, the independent variables are significantly intercorrelated; however, the extent of the correlation raises to only .28, far short of the problem caused when the correlations among independent variables approach +1.00 or −1.00. Furthermore, Howell (1987) suggested that the relative importance of each independent variable is indicated by the t statistic (indicated in Table 6 and all the multiple regression tables). It can be seen in our tables that relative importance is the same relative order as that of beta weights. Hence, what is said about the primary importance of DIT scores as one of the independent variables still stands in view of the t-test results.

Hence, one of the findings of the present study is the replication that the DIT is of first importance among independent variables (has higher beta weights and higher t-test scores). In all four replications (two in Narvaez et al., 1999, and both DIT1 and DIT2 results in the present study), this was the stable result; therefore, there does not seem to be a multicollinear problem in the stability of these results regarding beta weights.

It is true that the Rs in Narvaez et al. (1999) were generally in the range of .7 to .8. In the present study, Rs were in the .5 to .6 range. The difference in R may be due to

Table 8
Differences Between Students in Present Sample and Students in Narvaez et al. (1999), Sample 2

Variable	Present sample (n = 154) M	Present sample (n = 154) SD	Narvaez et al. (1999), Study 2 (n = 62) M	Narvaez et al. (1999), Study 2 (n = 62) SD	Difference (t test, df = 214)
DIT1-P	37.86	17.19	48.58	15.13	4.53****
POLCON	3.16	0.92	2.85	0.94	2.21*
FUNDA	58.13	12.81	55.48	14.78	1.24
ATHRI	145.93	15.18	159.16	17.36	5.28****

Note. DIT1-P = Defining Issues Test (original version), using P index; POLCON = political identity as conservative; FUNDA = religious fundamentalism; ATHRI = Attitudes Toward Human Rights Inventory (Getz, 1985).
*$p < .05$. ****$p < .0001$.

the peculiarities of the samples. As shown in Table 8, the student sample in the present study is generally more conservative, that is, lower in moral judgment, lower on advocacy for human rights, and more politically conservative than the student sample of Narvaez et al. (1999, Study 2). Future research may clarify whether views on public policy issues (ATHRI) are better predicted in more liberal groups (Narvaez et al., 1999, Study 2) than in more conservative groups (present study).

Conclusions

The four parts of this article indicate the following conclusions:

1. After 25 years of research using DIT1-P, there may now be a better DIT that is shorter, more updated, purges fewer participants, and has significantly better validity characteristics.

2. To the extent that DIT2 shows an improvement in validity trends over DIT1 in this study, the increase in validity seems to be attributable to the new ways of analyzing data (in indexing and in checking participant reliability) and not to the new dilemmas or new wording. We were expecting significant gains in validity for new dilemmas and wording (DIT1 vs. DIT2), for N2 versus P, and for new checks versus standard checks. Instead, we found significant improvements for the new analyses (N2 and new checks) but not for DIT1 over DIT2. Still, the practical advantages of DIT2 (i.e., it is shorter and updated and thus purges slightly fewer participants) recommend experimentation.

3. The reason that new checks show stronger trends on the validity criteria seems to be because they retain a wider range of scores, resulting in a fuller distribution of scores.

4. The present study supports the particular interpretation of Narvaez et al. (1999) regarding the combination and interaction of moral judgment with cultural ideology in the formation of opinions on public policy issues. DIT2 seems to operate in a way similar to DIT1 when used to predict attitudes toward public policy issues. More generally, this supports our view that Kohlbergian theories of morality are more useful in describing macromorality than micromorality.

Despite the long tradition in using the same dilemmas in Kohlbergian (1976, 1984) research (e.g., Heinz and the drug), this study suggests that there is nothing exceptional or magical about the DIT1's dilemmas and items, or about the classic Kohlberg dilemmas. It is possible to update, shorten, and revise the DIT without sacrificing validity. This should be encouraging for experimentation with new dilemmas and items. For instance, profession-specific dilemmas may be devised for a profession (e.g., for dentists, accountants, or teachers) in the hope of accounting better for profession-specific behavior.

This study reconfirms several basic findings about the moral judgment construct. First, the developmental, age and education trends are reconfirmed with DIT2 (i.e., moral judgment scores increase as age and education increases). Second, moral judgment scores are highly related to views on controversial public policy issues, as assessed by the ATHRI. Further, in multiple regression, moral judgment along with political identity and religious fundamentalism predict the ATHRI scores in combination more strongly than each independent variable alone, but each does not reduce to the other. This is consistent with the view expressed in Narvaez et al. (1999) about the relation of moral judgment to cultural ideology. Third, the new index, N2, as reported in Rest, Thoma, Narvaez, et al. (1997) shows advantages over the traditional ways of performing these calculations. Although there seems to be some gain in the power of trends using these new forms of analysis (N2 and new checks), the computations have become so labor intensive that hand scoring is no longer an option with N2 or new checks. To these replications, we add that DIT1 is highly correlated with DIT2 ($r = .79$) and that the 11 stories of DIT1 plus DIT2 show a very high degree of internal consistency (Cronbach's alpha = .90).

What are the practical implications of the present study? The findings encourage researchers to substitute DIT2 for DIT1. However, because this is only the first study with DIT2, with 200 participants—and because hundreds of studies have used the DIT1, involving about half a million participants—the older version must be regarded as the more established entity. Researchers for new projects must decide whether an updated, shorter, and slightly more powerful DIT2 with a short track record is preferable to the dated, longer, but better established DIT1. In any case, whether using DIT1 or DIT2, the new analyses (with N2 and new checks) should be employed. (Users of DIT1 can send previously scored data for rescoring, free of charge to the Center for the Study of Ethical Development.)

The most meaningful verdict on DIT2 must come from independent researchers beyond the site of development (Center for the Study of Ethical Development). The generalizability of DIT2, N2, and new checks must come from other researchers who may or may not find these innovations useful.

References

Bargh, J. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. Uleman & J. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York: Guilford Press.

Brown, D. G., & Lowe, W. L. (1951). Religious beliefs and personality characteristics of college students. *Journal of Social Psychology, 33,* 103–129.

Colby, A., Kohlberg, L., Speicher, B., Hewer, A., Candee, D., Gibbs, J., & Power, C. (1987). *The measurement of moral judgment* (Vols. 1–2). New York: Cambridge University Press.

Emler, N., Resnick, S., & Malone, B. (1983). The relationship between moral reasoning and political orientation. *Journal of Personality and Social Psychology, 45,* 1073–1080.

Ericsson, K. A., & Smith, J. (1991). *Toward a general theory of expertise.* New York: Cambridge University Press.

Getz, I. (1984). The relation of moral reasoning and religion: A review of the literature. *Counseling and Values, 28,* 94–116.

Getz, I. (1985). *The relation of moral and religious ideology to human rights.* Unpublished doctoral dissertation, University of Minnesota.

Gilligan, C. (1982). *In a different voice.* Cambridge, MA: Harvard University Press.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Holyoak, K. J. (1994). Symbolic connectionism: Toward third-generation theories of expertise. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise* (pp. 301–336). New York: Cambridge University Press.

Howell, D. C. (1987). *Statistical methods for psychology* (2nd ed.). Boston: Duxbury.

Hunter, J. D. (1991). *Culture wars: The struggle to define America.* New York: Basic Books.

Juergensmeyer, M. (1993). *The new cold war?* Berkeley, CA: University of California Press.

Killen, M., & Hart, D. (Eds.) (1995). *Morality in everyday life.* New York: Cambridge University Press.

Kohlberg, L. (1976). Moral stages and moralization: The cognitive developmental approach. In T. Lickona (Ed.), *Moral development and behavior* (pp. 31–53). New York: Holt, Rinehart, & Winston.

Kohlberg, L. (1984). *Essays on moral development: The nature and validity of moral stages, Vol. 2.* San Francisco: Harper & Row.

Kohlberg, L., Boyd, D. R., & Levine, C. (1990). The return of Stage 6: Its principle and moral point of view. In T. Wren (Ed.), *The moral domain: Essays in the ongoing discussion between philosophy and the social sciences* (pp. 1151–1181). Cambridge, MA: The MIT Press.

Lewicki, P. (1986). *Non-conscious social information processing.* New York: Academic Press.

Marty, M. E., & Appleby, R. S. (1991). *Fundamentalism observed.* Chicago: University of Chicago Press.

Marty, M. E., & Appleby, R. S. (1993). *Fundamentalism and the state.* Chicago: University of Chicago Press.

McClosky, H., & Brill, A. (1983). *Dimensions of tolerance: What Americans believe about civil liberties.* New York: Sage.

Narvaez, D. (1998). The influence of moral schemas on the reconstruction of moral narratives in eighth graders and college students. *Journal of Educational Psychology, 90,* 13–24.

Narvaez, D., Getz, I., Thoma, S. J., & Rest, J. (1999). Individual moral judgment and cultural ideologies. *Developmental Psychology, 35,* 478–488.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Pascarella, E. T., & Terenzini, P. (1991). *How college affects students: Findings and insights from twenty years of research.* San Francisco: Jossey-Bass.

Rawls, J. A. (1971). *A theory of justice.* Cambridge, MA: Harvard University Press.

Rest, J. (1979). *Development in judging moral issues.* Minneapolis: University of Minnesota Press.

Rest, J. (1986). *Moral development: Advances in research and theory.* New York: Praeger.

Rest, J., Cooper, D., Coder, R., Masanz, J., & Anderson, D. (1974). Judging the important issues in moral dilemmas—an objective test of development. *Developmental Psychology, 10,* 491–501.

Rest, J., & Narvaez, D. (Eds.). (1994). *Moral development in the professions: Psychology and applied ethics.* Hillsdale, NJ: Erlbaum.

Rest, J., & Narvaez, D. (1998). *Guide for DIT-2.* Unpublished manuscript. (Available from Center for Study of Ethical Development, University of Minnesota, 206 Burton Hall, 178 Pillsbury Dr., Minneapolis, MN 55455.)

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: A neo-Kohlbergian approach.* Mahwah, NJ: Erlbaum.

Rest, J., Narvaez, D., Mitchell, C., & Thoma, S. J. (1998a). *Exploring moral judgment: A technical manual for the Defining Issues Test.* Unpublished manuscript. (Available from the Center for the Study of Ethical Development, University of Minnesota, 206 Burton Hall, 178 Pillbury Dr., Minneapolis, MN 55455.)

Rest, J., Narvaez, D., Mitchell, C., & Thoma, S. J. (1998b). *How Test Length Affects Validity and Reliability of the Defining Issues Test.* Manuscript submitted for publication.

Rest, J., Thoma, S. J., & Edwards, L. (1997). Designing and validating a measure of moral judgment: Stage preference and stage consistency approaches. *Journal of Educational Psychology, 89,* 5–28.

Rest, J., Thoma, S. J., Narvaez, D., & Bebeau, M. J. (1997). Alchemy and beyond: Indexing the Defining Issues Test. *Journal of Educational Psychology, 89,* 498–507.

Schacter, D. L. (1996). *Searching for memory.* New York: Basic Books.

Snarey, J. (1985). The cross-cultural universality of social–moral development. *Psychological Bulletin, 97,* 202–232.

Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario Symposium* (Vol. 1, pp. 89–134). Hillsdale, NJ: Erlbaum.

Thoma, S., Narvaez, D., Rest, J., & Derryberry, P. (in press). The distinctiveness of moral judgment. *Educational Psychology Review.*

Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology; Learning, Memory, and Cognition, 8,* 336–342.

Uleman, J. S., & Bargh, J. A. (1989). *Unintended thought.* New York: Guilford Press.

Youniss, J., & Yates, M. (in press). Youth service and moral identity: A case for everyday morality. *Educational Psychology Review.*

Appendix

Sample Story From DIT2: The Famine

The small village in northern India has experienced shortages of food before, but this year's famine is worse than ever. Some families are even trying to feed themselves by making soup from tree bark. Mustaq Singh's family is near starvation. He has heard that a rich man in his village has supplies of food stored away and is hoarding food while its price goes higher so that he can sell the food later at a huge profit. Mustaq is desperate and thinks about stealing some food from the rich man's warehouse. The small amount of food that he needs for his family probably wouldn't even be missed.

What should Mustaq Singh do? Do you favor the action of taking the food? (*Check one*)

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Strongly favor	Favor	Slightly favor	Neutral	Slightly disfavor	Disfavor	Strongly disfavor

Rate the following issues in terms of importance (1 = great, 2 = much,

3 = some, 4 = little, 5 = no). Please put a number from 1 to 5 alongside every item.

1. ☐ Is Mustaq Singh courageous enough to risk getting caught for stealing?
2. ☐ Isn't it only natural for a loving father to care so much for his family that he would steal?
3. ☐ Shouldn't the community's laws be upheld?
4. ☐ Does Mustaq Singh know a good recipe for preparing soup from tree bark?
5. ☐ Does the rich man have any legal right to store food when other people are starving?
6. ☐ Is the motive of Mustaq Singh to steal for himself or to steal for his family?
7. ☐ What values are going to be the basis for social cooperation?
8. ☐ Is the epitome of eating reconcilable with the culpability of stealing?
9. ☐ Does the rich man deserve to be robbed for being so greedy?
10. ☐ Isn't private property an institution to enable the rich to exploit the poor?
11. ☐ Would stealing bring about more total good for everybody concerned or not?
12. ☐ Are laws getting in the way of the most basic claim of any member of a society?

Which of these 12 issues is the 1st most important? (*write in the number of the item*) ☐

Which of these 12 issues is the 2nd most important? ☐

Which of these 12 issues is the 3rd most important? ☐

Which of these 12 issues is the 4th most important? ☐

Note. An information package can be obtained from the Center for the Study of Ethical Development, University of Minnesota, 206 Burton Hall, 178 Pillsbury Drive Southeast, Minneapolis, Minnesota 55455. Electronic mail may be sent to narvaez@tc.umn.edu, or call (612) 624-0876.

Received November 10, 1998
Revision received February 16, 1999
Accepted February 16, 1999 ■