Administering Defining Issues Test Online: Do Response Modes Matter?

Yuejin Xu
Tarleton State University – Central Texas

Asghar Iran-Nejad and Stephen J. Thoma
The University of Alabama

Abstract

The purpose of the study was to determine comparability of an online version to the original paper-pencil version of Defining Issues Test 2 (DIT2). This study employed methods from both Classical Test Theory (CTT) and Item Response Theory (IRT). Findings from CTT analyses supported the reliability and discriminant validity of both versions. Findings from IRT analyses confirmed that both versions had comparable power of test-of-fit to the Rasch model. However, IRT analyses found that there were some variations in item difficulties and the patterns of item functions between the two versions. The study also examined students' satisfaction of DIT2-taking experience in the two survey response modes. The ANOVA results indicated that the online version of DIT2 was comparable to the paper-pencil version in terms of ease of use.

Introduction

Defining Issues Test (DIT) is "the most widely used measure of moral judgment development" (Thoma, 2002, p. 225). This test, sometimes referred to as DIT1, was designed by James Rest in 1974. A revised version (DIT2) was made available in 1999 (Rest, Narvaez, Thoma, & Bebeau, 1999). DIT2 makes several important changes including updating dilemmas and items, designing new indices, and producing new reliability checks to detect unreliable participants. However, both versions of the test have only used the traditional paper-and-pencil format up to this point.

With the development of modern technology, especially, the web technology, putting a survey online and collecting data by web HTML has been a popular practice (Couper, 2000; Hardre, et al., in press). Technically, it is not difficult to design an online version of DIT2 using web HTML. However, will the change of response mode from paper-and-pencil to online affect the original psychometric properties of DIT2? According to Cronbach (1990), "the conventional and computer versions of a test do usually measure the same variables, but difficulty or reliability can easily change. Whether the computer version is 'the same test' must be queried with each instrument in turn psychologically" (p. 48). Therefore, it is essential to empirically assess whether the paper-pencil version of DIT2 and the online version of DIT2 are equivalent.

Current methods for assessing equivalence fall into two categories: (1) classical test theory and common factor analysis, and (2) item response theory (IRT) procedures (Ferrando & Lorenzo-Seva, 2005). The classical test theory methods usually compare the estimated reliability indices (i.e., Cronbach alphas) of both versions (Carlbring et al., 2005; Pettit, 2002), the correlation coefficients to external criteria of both versions (Bunz, 2005; Im et. al, 2005), and the factor structures underneath both versions (Buchanan & Smith, 1999). Ferrando and Lorenzo-

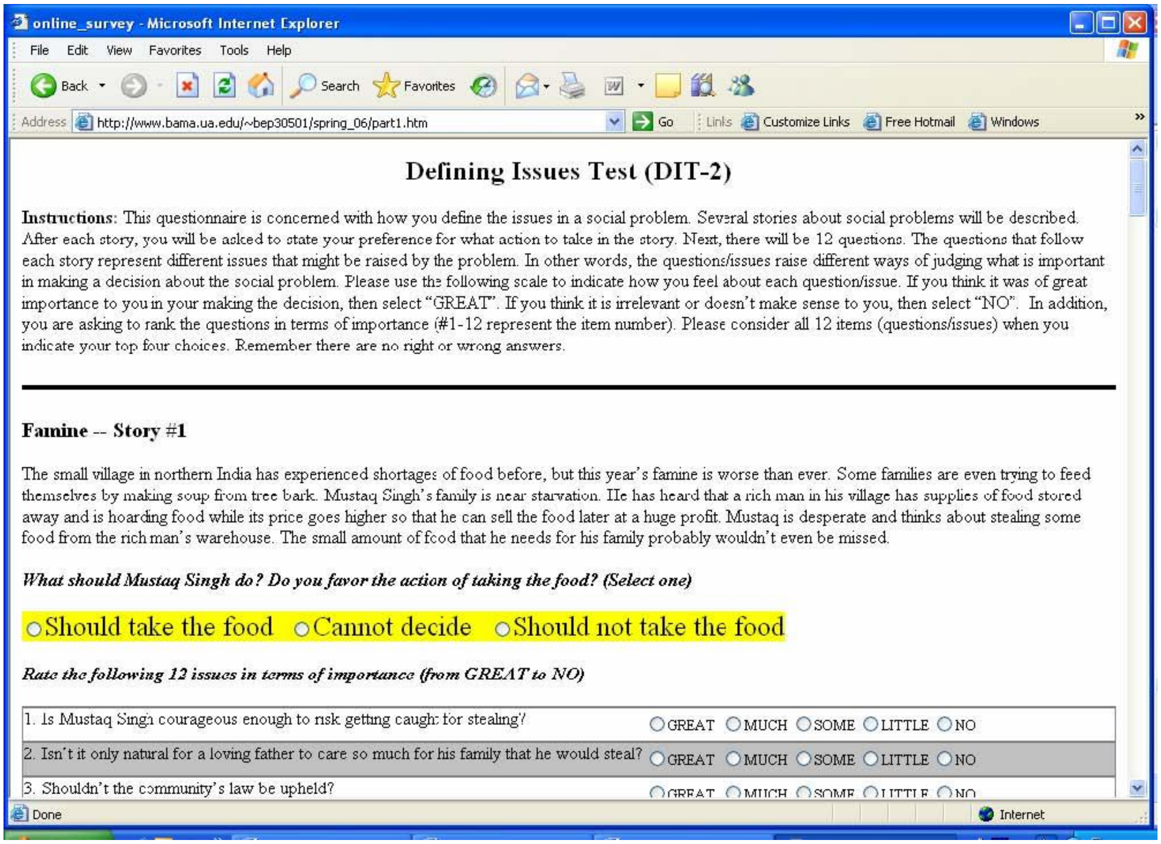
Seva (2005) summarized IRT-based procedures into "(a) correlating the trait estimates in both versions and correcting for attenuation; (b) comparing the information curves of both versions, and (c) assessing the linearity of the relations between the item difficulty and discrimination parameters estimated from both versions" (p. 194). This study employs methods from both classical test theory and IRT.

The Design of Online DIT2

Online survey has been used in many fields such as public opinions, and human-computer interaction. Over the years, there has been a growing literature on the design of online surveys (Couper, 2000, 2001; DeRouvray & Couper, 2002; Lazar & Preece, 1999) and on the cognitive processes in taking online surveys (Norman, 1991; Schwarz, 1999). A number of studies have also been conducted to compare the online survey method with other survey response modes (Couper, Blair, & Triplett, 1999; Dillman, Tortora, Conradt, & Bowker, 1998; Yun & Trumbo, 2000). Yun and Trumbo (2000) found that differences existed in the response characteristics of paper-and-pencil, e-mail, and web forms. They also reported no significant influences of survey mode in their substantive analyses.

MacElroy (1999) reported seven forms of online surveying: (1) e-mail (text), (2) bulletin boards, (3) web HTML, (4) web fixed-form interactive, (5) web customized interactive, (6) downloadable surveys, and (7) web-moderated interviewing. Of the seven forms, web HTML, or flat HTML survey form, seemed to be the most common form of online surveying. It requires less programming, offers more flexibility, operates faster, and costs less compared to other more advanced forms of online surveying. Consequently, the web HTML method was selected to design an online version of DIT2, and the principles for designing web questionnaires (Dillman, Tortora, & Bowker, 1998) were closely observed. The main design features of the online DIT2 were the following:

- The online version of DIT2 begins with a "Welcome" screen that instructs respondents on the action needed for proceeding to the next page.
- A consistent, standardized format scheme (font, font size, color, and spacing) is used to help respondents to answer the questions.
- By clicking "submit," respondents are automatically redirected to a thank-you webpage.
- A long, single page is used on which the respondent clicks radio buttons, fills in text boxes, and finally submits the information all at once. The page retains all the content of the paper-and-pencil version of DIT2 and all respondents need to do is to scroll down the screen to complete this survey. There are no pages to turn. Figure 1 was a screen shot of the DIT2 online version.


online_survey - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.bama.ua.edu/~bep30501/spring_06/part1.htm Go Links Customize Links Free Hotmail Windows
Done
Internet

Defining Issues Test (DIT-2)

Instructions: This questionnaire is concerned with how you define the issues in a social problem. Several stories about social problems will be described. After each story, you will be asked to state your preference for what action to take in the story. Next, there will be 12 questions. The questions that follow each story represent different issues that might be raised by the problem. In other words, the questions/issues raise different ways of judging what is important in making a decision about the social problem. Please use the following scale to indicate how you feel about each question/issue. If you think it was of great importance to you in your making the decision, then select "GREAT". If you think it is irrelevant or doesn't make sense to you, then select "NO". In addition, you are asking to rank the questions in terms of importance (#1-12 represent the item number). Please consider all 12 items (questions/issues) when you indicate your top four choices. Remember there are no right or wrong answers.

Famine -- Story #1

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? (Select one)

○ Should take the food ○ Cannot decide ○ Should not take the food

Rate the following 12 issues in terms of importance (from GREAT to NO)

1. Is Mustaq Singh courageous enough to risk getting caught for stealing?	○ GREAT ○ MUCH ○ SOME ○ LITTLE ○ NO
2. Isn't it only natural for a loving father to care so much for his family that he would steal?	○ GREAT ○ MUCH ○ SOME ○ LITTLE ○ NO
3. Shouldn't the community's law be upheld?	○ GREAT ○ MUCH ○ SOME ○ LITTLE ○ NO

Figure 1. A screen shot example of the online version of DIT2.

The primary purpose of this study was to determine comparability of an online version of DIT2 to the original paper-pencil version. Specifically, the study addressed three questions.
Research Question 1: Is the online version equivalent to the paper-pencil version of DIT2 in reliability indices?
Research Question 2: Is the online version comparable to the paper-pencil version in validity?
Research Question 3: Does survey mode affect respondents' degree of satisfaction in taking DIT2?

Methods

Participants and Setting

Respondents were 109 undergraduate and graduate students at a small teaching university in the Southwest. They were recruited from seven courses in the psychology and counseling program (undergraduate: human lifespan, abnormal psychology, research methods, elementary statistics; graduate: behavioral statistics, human development, and assessment and evaluation) in the beginning of the spring 2006 semester. A clustering sampling method was used. Out of the seven classrooms, three were randomly assigned to the paper-pencil DIT2-taking condition (a total of 47 students), and four were assigned to the online DIT2-taking condition (a total of 62 students).

Measures

Demographic variables: Participants were asked to report their age, sex, and education level. They reported their political view in a 5-point Likert scale with 1 standing for “very liberal,” and 5 for “very conservative.” For the age variable, this sample had a mean age of 32.78 ranging from 21 to 58. Most of the participants (95) were female (87.2%). There were 39 sophomores (36%) and 39 seniors (36%). The distribution of their age, gender, education level, and political view are reported in Table 1. There was no major difference in distribution of gender and political view across the two conditions: paper-pencil vs. online DIT2. However, the participants in the online condition turned out to be slightly younger and lower in educational level than the participants in the paper-pencil condition.

Table 1: *Frequency (and Percentage) of Age, Gender, Education Level, and Political View in the Paper-pencil and Online Conditions*

Variables	Overall (%)	Paper-pencil (%)	Online (%)
Age			
20-29	46 (43.4%)	15 (34.1%)	31 (50%)
30-39	31 (29.2%)	14 (31.8%)	17 (27.4%)
40-49	22 (20.8%)	10 (22.7%)	12 (19.4%)
50-59	7 (6.6%)	5 (11.4%)	2 (3.2%)
Gender			
Female	95 (87.2%)	41 (87.2%)	54 (87.1%)
Male	14 (12.8%)	6 (12.8%)	8 (12.9%)
Education level			
Voc/Tech	1 (.9%)	0	1 (1.6%)
Jr. college	1 (.9%)	1 (2.1%)	0
Freshman	21 (19.3%)	0	21 (33.9%)
Sophomore	39 (35.8%)	15 (31.9%)	24 (38.7%)
Junior	6 (5.5%)	3 (6.4%)	3 (4.8%)
Senior	39 (35.8%)	27 (57.4%)	12 (19.4%)
Prof. degree	2 (1.8%)	1 (2.1%)	1 (1.6%)
Political view			
Very Lib	11 (10.1%)	3 (6.4%)	8 (12.9%)
Somewhat lib	33 (30.3%)	16 (34%)	17(30.3%)
Neither	31 (28.4%)	11 (23.4%)	20 (32.3%)
Somewhat con	25 (22.9%)	11 (23.4%)	14 (22.6%)
Very con	9 (8.3%)	6 (12.8%)	3 (8.3%)

Defining Issues Test 2 (DIT2) (Rest et al., 1999) is an updated version of the original DIT (Rest, 1979). The original DIT has demonstrated robust validity and reliability in hundreds of studies. The Cronbach alpha of the DIT is in the upper .70s/ low .80s. Test-retest reliability is about the same. The DIT is sensitive to moral education interventions. It is also significantly linked to many pro-social behaviors and to desired professional decision making. Defining Issues Test 2 retains the psychometric properties of the original DIT and improves on validity. Defining Issues Test 2 is a paper-pencil measure of moral judgment derived from Kohlberg’s theory. An online version was designed and used together with its original paper-pencil version. DIT2 consists of five dilemma stories. Participants first choose one of three listed courses of action that follow from each story. Next, they rate the level of importance of their decision by responding to

12 statements on a Likert-type scale (1= No importance, 2= Little importance, 3= Somewhat important, 4= Much importance, 5= Great importance). Finally, participants rank the 12 statements in terms of importance and list their top four picks. Two main developmental indices provided by DIT2 are the P score, which is the principled score, and the N2 score, which is the P score adjusted for lower stage reasoning. The N2 score has been reported to be equivalent to P score but generally outperforms the P score in construct validity on six criteria (Rest, Thoma, Narveaz, & Bebeau, 1997). A high P-sore/ N2 score indicates high-level moral reasoning and decision making. In this study, both P score and N2 score were examined. DIT2-taking experience was measured by a 7-point Likert scale with 1 standing for “not at all enjoyable,” 4 standing for “somewhat enjoyable,” and 7 standing for “extremely enjoyable.”

Computer technology knowledge was assessed by two indices. One were their self-reported ratings of internet skills on a 5-point Likert scale with 1 standing for “not at all,” 2 “not very skilled,” 3 “fairly skilled,” 4 “very skilled,” and 5 “expert.” The other index rated their understanding on 15 technical terms/concepts (for example: HTML) on a 5-point Likert scale with 1 standing for “none,” 3 “somewhat understanding,” and 5 “full understanding.” Both indices were adapted from a digital literacy measure by Hargittai (2005).

A self-designed survey was used to assess participants’ general attitudes toward computer technology. It consisted of 10 statements. One sample item was “I like to use computers.” Respondents were asked to use a 5-point Likert scale to indicate how they feel about each statement with 1 standing for “not at all true,” and 5 standing for “very true.” The Cronbach alpha for this scale is .86, indicating a reasonable level of reliability.

Procedures

Participation was solicited from students enrolled in seven courses. Both Institutional Review Board (IRB) approval of the use of human subjects and instructors’ approval were first obtained. A regular class meeting time was scheduled for this project. In the paper-pencil condition, willing participants signed a consent form and completed the DIT2 and a survey package including their DIT2-taking experience, computer knowledge, and attitude. In the online condition, willing participants were given a detailed instruction sheet to do the same as those in the paper-pencil condition on the website, at their own convenience, within the given time frame. The website was closed when the data was collected.

The data collected in the paper-pencil condition were entered into SPSS whereas data collected in the online condition were imported into SPSS.

Data Analysis

Research Question 1: Is the online version equivalent to the paper-pencil version of DIT2 in reliability indices?

In *Guide for DIT2*, Bebeau and Thoma (2003) recommended using the story score as the basic unit of internal reliability. “The 5 stories – not the items – are used as the units for calculating reliability because ranking data is ipsative (that is, if one item is ranked in first place, then no other item can be ranked in first place)” (Bebeau & Thoma, 2003, p. 30). Cronbach alphas were calculated for each version of DIT2 (paper-pencil and online) using both the N2 index and the P index at the story level. The Rasch model was selected for IRT analysis. Data requirements for Rasch analysis were discussed. Rasch model requires dichotomous or polytomous data. Therefore, P index was used in IRT analysis. The Rasch Unidimensional Measurement Models 2010 (RUMM 2010) computer program was used to calculate separation index, one reliability index in IRT analysis, for each version of DIT2.

Research Question 2: Is the online version comparable to the paper-pencil version in validity?

The RUMM 2010 program was used for the IRT analysis to calibrate item difficulties (each story in DIT2) and person ability (the latent variable of morality) estimates. Specifically, the item-person map, item functions, and fit statistics for the scale were examined to test internal construct validity of both versions of DIT2.

Rest, Thoma, and Edwards (1997) described 7 criteria for establishing the construct validity of DIT2. Due to the limit of this sample (the sample is not fully randomized and the sample did not cover the full range of educational levels), we mainly examine discriminant validity of the two versions of DIT2. Correlations between moral judgment scores and DIT2-taking experience index, respondents' technology knowledge, and attitudes toward technology were computed to determine the discriminant validity in the two conditions.

Research Question 3: Does survey mode affect respondents' degree of satisfaction in taking DIT2?

A one-way analysis of variance (ANOVA) was conducted to examine whether survey mode (paper-pencil versus online) affects respondents' DIT2-taking experience. A correlation matrix was designed for each condition (paper-pencil and online) to demonstrate the correlations among DIT2-taking experience index and respondents' technology knowledge and attitudes toward technology.

Results

Research Question 1: Is the online version equivalent to the paper-pencil version of DIT2 in reliability indices?

Besides moral reasoning developmental indices (P score and N2 score), the scoring of DIT2 also provides useful reliability checks. The DIT2 reliability check indices were closely examined. Of the 109 participants, 7 failed to pass the reliability checks (3 from the paper-pencil condition and 4 from the online condition). Students who failed to pass the reliability checks were excluded from further analysis.

Table 2 shows the mean and standard deviation of N2 score for each story in the paper-pencil condition and the online condition. Participants in the paper-pencil condition scored higher in N2 score for each story than participants in the online condition. This is more likely related to the difference in distribution of age and level of education in that the participants in the paper-pencil condition happened to have a higher level of education. Education level has been recognized as a major contributor to moral judgment (Bebeau & Thoma, 2003). The correlation between each story N2 and the N2 score for the whole scale (item-to-total correlation) and reliability (Cronbach alphas) are presented in Table 3. Both versions of DIT2 demonstrate good item-to-total correlation and show comparable internal consistence (in the paper-pencil condition, α = .649; in the online condition, α = .700).

Table 2: *Descriptive Statistics for the Paper-pencil and the Online Version of DIT2*

	Paper-pencil (N=44)		Online (N=58)	
Item/Story	Mean	SD	Mean	SD
1. Famine	2.39	4.23	1.32	3.56
2. Reporter	4.90	4.14	2.33	4.19
3. School Board	5.82	4.12	4.01	4.22
4. Cancer	5.35	4.66	3.90	4.15
5. Demonstration	2.89	4.77	1.67	4.12

Table 3: *Item-to-total Correlation Coefficients and Reliability Coefficients for the Paper-pencil and the Online Version of DIT2 Using N2 score*

	Total N2 score	
Item/Story	Paper-pencil (N=44)	Online (N=58)
1. Famine	.603**	.631**
2. Reporter	.482**	.681**
3. School Board	.763**	.637**
4. Cancer	.600**	.694**
5. Demonstration	.771**	.728**
Reliability coefficients	.649	.700

**Correlation is significant at the 0.01 level (2-tailed).

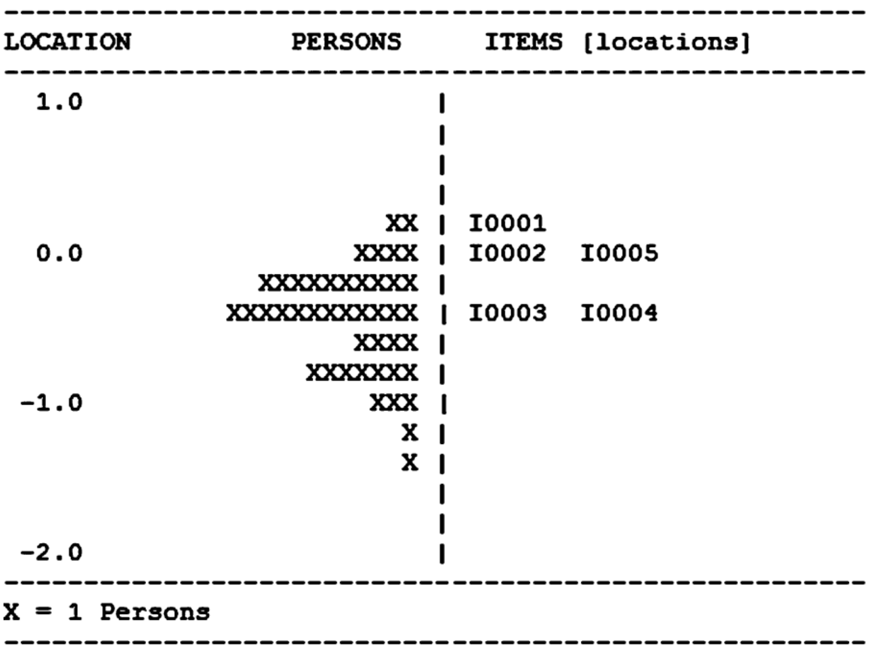
Item response theory (IRT) extends the concept of reliability by including the person ability factor. In IRT, reliability refers to the degree of precision at different values of person ability (theta). Of the many models in IRT, the Rasch model was selected for the IRT analyses in this study. Rasch model is a unidimensional model which asserts that the easier the item, the more likely it will be endorsed, and the more able the person, the more likely the person will endorse a difficult item compared with a less able person. The one parameter Rasch model assumes that the probability of a person to endorse a given item is a logistic function of the relative distance between the item location and the person location. A Rasch model IRT analysis requires that the item data be in either dichotomous or polytomous format (RUMM, 2000). P scores for each story were used as the item data for IRT analysis. "A P score is calculated on the basis of ranking data. If a participant ranks a principled item as 'most important,' then this increases the P score by 4 points (in second place, by 3 points; in third place, by 2 points; in fourth place, by 1 point)" (Rest, Thoma, Narvaez, & Bebeau, 1997, p. 500). The P score at the story level could range from 0 to 10. The P score for the whole scale is the total number of points across the 5 stories and is converted to a percentage. P scores can range from 0 to 95. Compared to the N2 scores, P scores is in polytomous format, more suitable for IRT analyses.

Separation index is one of the reliability indices yielded by RUMM 2010 program. Separation index, or the person separation index "depends in part on the actual variance of the persons. It tells whether person with higher locations than persons with lower locations tend to obtain higher scores on items or not" (RUMM 2020, p. 9). In the paper-pencil condition, the separation index is .561, whereas in the online condition, the separation index is .531. Both indicate reasonable separation of items along the theoretical construct of moral judgment. The Cronbach alphas based on the P scores are also generated by RUMM 2010. The two versions of DIT2 have almost equivalent Cronbach alphas based on the P scores (in the paper-pencil condition, α = .562; in the online condition, α = .536).

Research Question 2: Is the online version comparable to the paper-pencil version in validity?

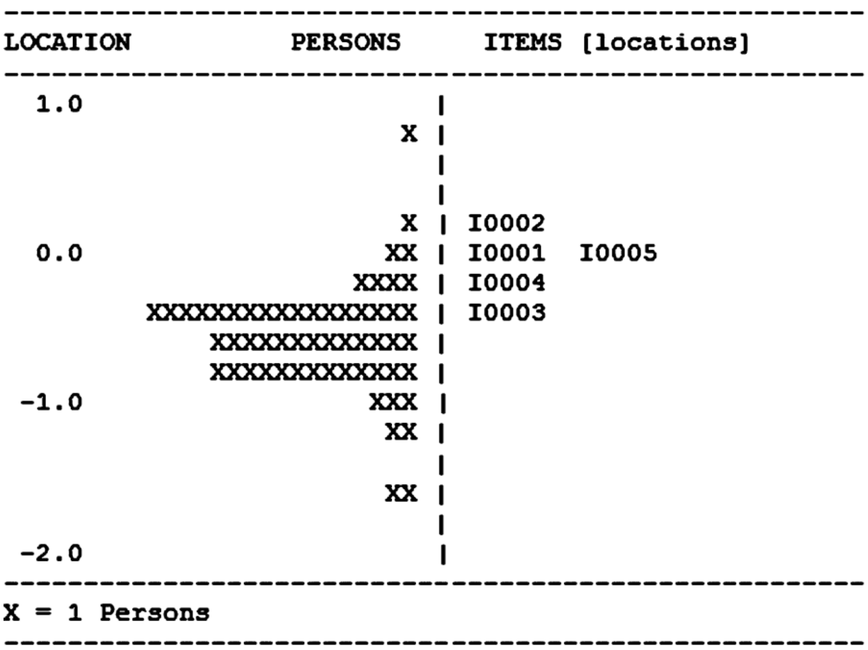
The Rasch model is used to examine the internal construct validity of the paper-pencil and the online version of DIT2, and to further examine whether the items are working together in a way that is explained by the model. The RUMM-generated IRT results are reported in the following order: item-person map, individual item-fit statistics, and overall test-of-fit statistics.

Item-person map depicts the locations of persons and items on the continuums of ability and difficulty. Figure 2 and Figure 3 show the item-person map for the paper-pencil and the online condition respectively. Person ability is expressed as logit for probability of answering 0 (low p value, or low moral judgment score) to 10 (high p value, or high moral judgment score) for each story. In the item-person map for the paper-pencil condition (Figure 2), logits for person ability are approximately ranged from -1.8 to .2. Logits for item difficulty are ranged from -.28 to .2. The person ability index is similar to the item difficulty, indicating that many students may find it not easy to solve the moral dilemmas. Item I0003 (School Board) and item I0004 (Cancer) are relatively easy for most persons to endorse. There is one person for each “x” symbol on the map. In the item-person map for the online condition (Figure 3), logits for person ability range approximately from -1.5 to .9. Logits for item difficulty range from -.28 to .30. A close examination of Figure 3 revealed that only one person’s ability index is highly above the item difficulty, the rest were either similar or below the item difficulty index. Item-person map for the online condition also show that items I0003 (School Board), and I0004 (Cancer) are relatively easy for most persons to endorse.



```
--------------------------------------------------------
LOCATION          PERSONS     ITEMS [locations]
--------------------------------------------------------
  1.0                       |
                            |
                            |
                            |
                         XX | I0001
  0.0                  XXXX | I0002  I0005
                 XXXXXXXXXX |
               XXXXXXXXXXXX | I0003  I0004
                       XXXX |
                    XXXXXXX |
 -1.0                   XXX |
                          X |
                          X |
                            |
                            |
 -2.0                       |
--------------------------------------------------------
X = 1 Persons
--------------------------------------------------------
```

Figure 2. Item-person map of the paper-pencil version of DIT2.



```
----------------------------------------------------
LOCATION         PERSONS     ITEMS [locations]
----------------------------------------------------
  1.0                     |
                        X |
                          |
                          |
                        X | I0002
  0.0                  XX | I0001  I0005
                     XXXX | I0004
        XXXXXXXXXXXXXXXXX | I0003
            XXXXXXXXXXXXX |
            XXXXXXXXXXXXX |
 -1.0                 XXX |
                       XX |
                          |
                       XX |
                          |
 -2.0                     |
----------------------------------------------------
X = 1 Persons
----------------------------------------------------
```

Figure 3. Item-person map of the online version of DIT2.

Item-fit statistics for the two conditions are presented in Table 4 and Table 5. Each table includes type, location, standard error, residual, degrees of freedom (associated with residual), data points, chi-square, probability, and degrees of freedom (associated with chi-square and probability) for each item. Items are ordered by location. Item locations are logits for item difficulties, ranging from -4.00 to +4.00. A negative logit value of item location indicates that that item is easy to endorse whereas a positive logit value of item location indicates that item is difficult to endorse. Since the response format for the DIT2 is P score at the story level ranging from 0 to 10, item type used in this study is polytomous (or poly). The residual values represent the difference between the Rasch model's theoretical expectations and the actual performance. Degrees of freedom list the degrees of freedom associated with the residual value. Data points indicate how many participants respond to each item. In our case, there is no missing data. Chi-square is the item-trait interaction statistic for each item, which reflects the degree of invariance of the trait, in this case, moral judgment. A large, significant chi-square value usually means that the location of the item difficulty is not invariant across the trait. Prob refers to the probability of the occurrence of the chi-square value for the degree of freedom associated with the item-trait interaction of that item. For each item, we require residuals within the range -2.5 to 2.5 (which is also the default setting of RUMM 2010 program) and a nonsignificant chi-square value. Given that there were five items/dilemmas in the DIT2, we applied a Bonferroni adjustment to the chi-square probability value, giving a value of .002. Misfit of items can indicate a lack of unidimensionality, namely that a given item (moral dilemma) does not fit into the underlying construct of moral judgment.

In the paper-pencil condition (Table 4), all of the five items displayed residuals within the range -2.5 to +2.5. Moreover, all of the five items had nonsignificant chi-square values, indicating good fit to the Rasch model. Item I0004 (Cancer) has the lowest residual (-.011) suggesting that the actual performance is close to the model's expectation, whereas item I0002

(Reporter) has the largest residual (1.251) suggesting that the actual performance is somewhat different from the Rasch model's expectation. Item I0003 (School Board), and item I0004 (Cancer) are dilemmas with solutions in the easy range. Item I0001 (Famine) and item I0005 (Demonstration) fall into the difficult range.

In the online condition (Table 5), the five items also demonstrated good fit to the Rasch model. Item I0003 (School Board) has the lowest residual (-.025), whereas item I0001 (Famine) has the largest residual (.788). Item I0003 (School Board), and item I0004 (Cancer) are the two easiest items to solve. Item I0002 (Reporter) and item I0005 (Demonstration) are the two most difficult items to endorse.

Table 4: *Individual Item-fit Statistics for the Paper-pencil Version of DIT2*

Item/Story	Type	Location	SE	Residual	DF	DatPts	Chisq	Prob	DF
I0003 School Board	Poly	-.279	.08	-.145	31.40	44	4.355	.090	2
I0004 Cancer	Poly	-.273	.07	-.011	31.40	44	1.769	.397	2
I0002 Reporter	Poly	.114	.08	1.251	31.40	44	4.568	.078	2
I0005 Demonstration	Poly	.185	.08	.316	31.40	44	.850	.645	2
I0001 Famine	Poly	.252	.08	1.102	31.40	44	.489	.777	2

Table 5: *Individual Item-fit Statistics for the Online Version of DIT2*

Item/Story	Type	Location	SE	Residual	DF	DatPts	Chisq	Prob	DF
I0003 School Board	Poly	-.281	.06	-.025	42.60	58	.138	.931	2
I0004 Cancer	Poly	-.179	.06	.205	42.60	58	.264	.873	2
I0001 Famine	Poly	.051	.07	.788	42.60	58	2.645	.247	2
I0005 Demonstration	Poly	.105	.07	.245	42.60	58	.827	.652	2
I0002 Reporter	Poly	.304	.08	.101	42.60	58	3.287	.172	2

Overall test-of-fit summary statistics include item-person interaction, item-trait interaction, and power of test-of-fit. These analyses are summarized in Table 6 (for the paper-pencil condition) and Table 7 (for the online condition). The locations of the items have a mean of 0. This is the arbitrary constraint imposed on the location parameters of the items in RUMM 2010 in the estimation of the parameters. The standard deviation (SD) of the items (.256 in the paper-pencil condition; .233 in the online condition), is the empirically derived SD of the item locations. Fit residual of the items is a statistic that provides information on the fit of the data to the model from the perspective of the items. Therefore, "if the data accord with the model, its mean across all items should be close to 0 and its standard deviation close to 1" (RUMM 2020, p. 6). However, this may only be an approximation. In the paper-pencil condition, the values are respectively .503 and .640. In the online condition, the values are respectively .263 and .312. Ideally, the skewness and kurtosis values should be close to 0.

While the location of items has a priori constraint, the location of the persons is an empirical matter. In the paper-pencil condition, the mean = -.382 and the SD = .356. In the online condition, the mean = -.487 and the SD = .378. According to the Rasch model, the locations of the persons in both conditions are quite comparable. The mean and SD of fit residual across all persons should be close to 0 and 1 respectively. In the paper-pencil condition, these respective values are -.116 and 1.110. In the online condition, these values are -.246 and 1.100 respectively.

The item-trait interaction, as a chi-square statistic, reflects "the deviation from the model by groups of people defined by their ability level (in moral judgment) and requires a non-significant chi-square i.e. > .05" (Lundgren-Nilsson, et al, 2005, p. 24). In the paper-pencil condition, item-trait interaction was not significant (chi-square = 12.032, p = .264), suggesting that the paper-pencil version as a whole meets the Rasch model's expectation. In the online condition, total item chi-square is 7.161 and total degree of freedom is 10, which was not significant (chi-square = 7.161, p = .734), also meeting the Rasch model's expectation.

Power of test-of-fit is based on separation index. The person separation index plays an important role in interpreting the fit statistics in the Rasch model, for it indicates the actual variance of the person across the items. Both the paper-pencil and the online version of DIT2 demonstrate reasonable power of test-of-fit. In the paper-pencil condition, separation index is .561, whereas in the online condition, separation index is .531.

Table 6: *Test-of-Fit Summary Statistics for the Paper-pencil Version of DIT2*

Item-Person Interaction				
	Items		Persons	
	Location	Fit Residual	Location	Fit Residual
Mean	0.000	0.503	-0.382	-0.116
SD	0.256	0.640	0.356	1.110
Skewness		0.169		-2.226
Kurtosis		-2.177		7.903
Correlation		0.147		0.107
Complete data DF =	0.714			
Item-Trait Interaction				
Total Item Chi Square		12.032		
Total Degree of Freedom		10.000		
Total Chi Square Prob		0.264		
Reliability Indices				
Separation Index		0.561		
Cronbach Alpha		0.562		
Power of Test-of-Fit				
Power is REASONABLE [Based on SepIndex of 0.561]				

Table 7: *Test-of-Fit Summary Statistics for the Online Version of DIT2*

Item-Person Interaction				
	Items		Persons	
	Location	Fit Residual	Location	Fit Residual
Mean	0.000	0.263	-0.487	-0.246
SD	0.233	0.312	0.378	1.100
Skewness		0.771		-1.060
Kurtosis		-1.226		1.490
Correlation		0.026		-0.079
Complete data DF =	0.734			
Item-Trait Interaction				
Total Item Chi Square		7.161		
Total Degree of Freedom		10.000		
Total Chi Square Prob		0.702		
Reliability Indices				
Separation Index		0.531		
Cronbach Alpha		0.536		
Power of Test-of-Fit				
Power is REASONABLE				
[Based on SepIndex of 0.531]				

In order to further compare item difficulties between the paper-pencil and online versions of DIT2, the five stories were rank ordered from easy to difficult for each condition (Table 8). The two conditions did not share the exact order in item difficulties. Participants in the paper-pencil condition responded that item I0001 (Famine) was more difficult to endorse than item I0002 (Reporter). In contrast, participants in the online condition responded that item I0002 (Reporter) was more difficult to endorse than item I0001 (Famine).

Table 8: *Comparison of Item Difficulty between the Paper-pencil and the Online Version of DIT2*

	Paper-pencil			Online		
	Item/Story		Logit	Item/Story		Logit
Easy	I0003	School Board	-.279	I0003	School Board	-.281
	I0004	Cancer	-.273	I0004	Cancer	-.179
	I0002	Reporter	.114	I0001	Famine	.051
	I0005	Demonstration	.185	I0005	Demonstration	.105
Difficult	I0001	Famine	.252	I0002	Reporter	.304

Table 9 presents correlation coefficients between moral judgment scores (N2 score for the whole scale) and other constructs including DIT2-taking experience, computer technology knowledge, and attitude toward technology. The small and non-significant correlation coefficients indicate that the two versions of DIT2 are different from the other measures.

Table 9: *Discriminant Validities of the Paper-pencil and the Online Version of DIT2*

Discriminant validity	Paper-pencil	Online
Correlation coefficient between N2 score and DIT2-taking experience scale	-.012 (N=44)	.078 (N=56)
Correlation coefficient between N2 score and computer technology knowledge	.025 (N=44)	-.016 (N=58)
Correlation coefficient between N2 score and internet knowledge	-.062 (N=44)	-.101 (N=58)
Correlation coefficient between N2 score and attitude towards computer technology	-.047 (N=44)	-.209 (N=58)

Note. None of the correlation coefficient was significant.

Research Question 3: Does survey mode affect respondents' degree of satisfaction in taking DIT2?

The means and standard deviations of the DIT2-taking experience index for each condition (paper-pencil and online) are reported in Table 10. The mean scores for the DIT2-taking experience in both conditions are around 4, which is "somewhat enjoyable."

Table 10: *Means and Standard Deviations of DIT2-taking Experience Index in Two Conditions*

	M	SD	N
Paper-pencil	4.39	1.50	44
Online	4.09	1.49	56

A one-way analysis of variance was conducted to assess the effect of survey mode on students' DIT2-taking experience. The independent variable, survey mode, included two conditions: paper-pencil and online. The dependent variable was the self-reported DIT2-taking experience index. Levene's test of equality of error variance indicated that the error variance of the dependent variable was equal across groups: F (1, 98) = .179. p = .681, meeting the homogeneity of variance assumption for ANOVA. The ANOVA was not significant, F (1, 98) = .973, p > .05, indicating that the DIT2-taking experiences were not different for two conditions. The strength of relationship between test-taking condition and DIT2-taking experience indicator, as accessed by partial η^2, was small (partial η^2= .010). Some researchers advocate reporting post hoc power analysis when statistically non-significant results are found (Onwuegbuzie & Leech, 2004). Using the actual observed effect sizes pertaining to the difference between the paper-pencil and the online condition, the post hoc statistical power estimate was .164, which represents low statistical power for detecting the small observed effect size.

To further examine the relationships among DIT2-taking experience and respondents' technology knowledge and attitude, correlation matrixes were designed for the paper-pencil condition (Table 11), and the online condition (Table 12).

Table 11: *Correlation Matrix for the Paper-pencil Condition*

	1	2	3	4
1. DIT2-taking experience	1			
2. Technology knowledge	.213	1		
3. Internet ratings	.051	.599**	1	
4. Attitudes toward technology	.248	.642**	.801**	1

**. Correlation is significant at the .01 level (2-tailed).

Table 12: *Correlation Matrix for the Online Condition*

	1	2	3	4
1. DIT2-taking experience	1			
2. Technology knowledge	.403**	1		
3. Internet ratings	.396**	.574**	1	
4. Attitudes toward technology	.341*	.561**	.704**	1

*. Correlation is significant at the .05 level (2-tailed).
**. Correlation is significant at the .01 level (2-tailed).

Table 11 and 12 clearly show that participants' knowledge of computer and internet technology was positively related to their attitudes toward computer technology. In the online condition, participants' rating of their DIT2-taking experience was significantly correlated with their attitudes toward and knowledge of computer technology.

Discussion and Conclusions

Equivalence in Reliability of Two Versions of DIT2

Research Question 1 addressed the equivalence in reliability of two versions of DIT2 using approaches from both classical test theory and IRT. Our findings from procedures based on the classical test theory indicated that the online version was comparable to the paper-pencil version in terms of item-to-total correlation coefficients and reliability coefficients. Findings from IRT analysis further confirmed that the two versions had similar separation index and the same reasonable power of test-of-fit, meeting the unidimensionality requirement of the Rasch model. Moreover, according to *Guide for DIT2* (Bebeau & Thoma, 2003), Cronbach alpha for N2 at the story level was around .81. However, if the "sample does not contain the entire range of educational levels" (Bebeau & Thoma, 2003, p. 30), the Cronbach alpha is expected to be lower than .81. The Cronbach alpha for the online version (.700) and the paper-pencil version (. 649) fit into this description very well.

Equivalence in Disciminant Validity of Two Versions of DIT2

IRT analyses yielded mixed findings. On the one hand, both versions met the unidimensionality requirement of the Rasch model, and achieved a reasonable power of test-of-fit. Consequently, both versions demonstrated sufficient internal construct validity under IRT analysis. On the other hand, there existed some variations in item difficulties across the two versions. The patterns of item functions are not fully identical across the two versions. What was regarded relatively easy to endorse in one condition became most difficult to solve in the other condition. For example, participants in the paper-pencil condition responded that item I0002 (Reporter) was relatively easy to endorse. However, participants in the online condition found that item I0002 (Reporter) most difficult to endorse. This may be caused by the small sample size and the existence of outlier in the online condition.

Rest, Thoma, and Edwards (1997) described seven criteria for construct validity: “(1) differentiation of naturally occurring groups in terms of presumed expertise in moral judgment, (2) correlations of moral judgment with moral comprehension, (3) longitudinal change as a function of age and of enriching experiences, (4) sensitivity to moral education interventions, (5) links of moral judgment with behavior, (6) links of moral judgment with political attitudes, and (7) reliability” (p. 14). Due to the limits of the sample (the sample was not fully randomized and not covering the whole range of education levels) and the design (no inclusion of moral comprehension, behavior, intervention, or attitude measures), we can compare 1 (i.e., reliability) out of the 7 criteria of construct validity between the paper-pencil version and the online version of DIT2. However, correlation coefficients of moral judgment with DIT2-taking experience, technology knowledge, and attitude were low and non-significant in both conditions, indicating that both versions had comparable discriminant validity.

Survey Mode and DIT2-taking Experience

Our findings indicate that survey mode did not affect respondents’ ratings on their DIT2-taking experience. This implies that our online version of DIT2 was comparable to the paper-pencil version in terms of ease of use. However, the post hoc power analysis yielded a low power estimate. Replications are thus needed to verify the reliability of the present findings of no DIT2-taking experience differences in two different test-taking conditions. The correlation matrix for the online condition did show that people who had more positive attitude toward and more knowledge of the computer and internet technology enjoyed the online version of DIT2 more. Not surprisingly, this effect did not show in the paper-pencil version.

Implications and conclusions

In this study, psychometric properties of an online version and the original paper-pencil version of DIT2 were examined. The overall findings supported the reliability and discriminant validity of the paper-pencil and online versions of DIT2. Based on our discussions of the findings, we want to conclude the paper with the following implications.

First, our study expands existing literature on DIT2 by introducing the technology factor. With the development of internet technology, administering DIT2 online has the potential of making the use of DIT2 more convenient and flexible.

Second, computer knowledge and attitude do not affect moral judgment, however, they relate to the test-taking experience in the online condition. Therefore, it is necessary to find a better way in designing the online survey so as to make it accessible to those who may not know much about technology.

Third, this study is relevant to online learning not merely because it involves in using web HTML for an online survey to save time and resources in data collection, but because online learning implies, and may even require online assessment. Using methods from both Classical Test Theory (CTT) and Item Response Theory (IRT), the study expands our understanding of the comparability of an online version of DIT2 to the original paper-pencil version, and has special significance for the fields of online assessment.

References

Bebeau, M. J., & Thoma, S. J. (2003). *Guide for DIT2*. University of Minnesota.

Bunz, U. (2005). Using scantron versus an audience response system for survey research: Does methodology matter when measuring computer-mediated communication competence? *Computers in Human Behavior, 21*(2), 343-359.

Carlbring, P., Brunt, S., Bohman, S., Austin, D. Richards, J., Ost, L., & Andersson, G. (2005). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior, 22*(3), 545-553.

Couper, M. P. (2000). Web surveys: a review of issues and approaches. *Public Opinion Quarterly, 64*(4), 464-494.

Couper, M. P., Blair, J., & Triplett T. (1999). A comparison of mail and e-mail for a survey of employees in federal statistical agencies. *Journal of Official Statistics, 15*, 39-56.

Couper, M. P., Traugott, M., & Lamias, M. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*(2), 230-253.

DeRouvray, C., & Couper, M. P. (2002). Designing a strategy for reducing 'No Opinion' responses in Web-based surveys. *Social Science Computer Review SSCREH, 20*(1), 3-9.

Dillman, D. A., Tortora, R. D., & Bowker, D. (1998). *Principles for constructing web surveys.* SESRC Technical report 98-50, Pullman, Washington.

Dillman, D. A., Tortora, R. D., Conradt, J., & Bowker D. (1998). Influence of plan vs. fancy design on response rates of Web surveys. Presented at Joint Statistical Meetings, Dallas, Texas. August 1998. Retrieved December 2, 2002 from http://survey.sesrc.wsu.edu/dillman/papers/asa98ppr.pdf

Ferrando, P. J., & Lorenzo-Seva, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and internet-administered questionnaires. *Psychological Methods, 10*(2), 193-205.

Hardré, P. L, Crowson, H. M, Ly, C., & Xie, K. (in press). Testing differential effects of computer-based, web-based, and paper-based administration of questionnaire research instruments. *British Journal of Educational Technology.*

Hargittai, E. (2005). Survey measures of web-oriented digital literacy. *Social Science Computer Review, 23*(3), 371-379.

Im. E., Chee, W., Bender, M., Cheng, C., Tsai, H., Kang, N., & Lee, H. (2005). The psychometric properties of pen-and-pencil and internet versions of the midlife women's symptom index (MSI). *International Journal of Nursing Studies, 42*, 167-177.

Lazar, J., & Preece, J. (1999). Designing and implementing Web-based surveys. *Journal of computer information systems, 39*, 63-67.

Lundgren-Nilsson, A., Grimby, G., Ring, H., Tesio, L., Lawton, G., Slade, A., et al. (2005). Cross-cultural validity of functional independence measure items in stroke: A study using Rasch analysis. *Journal of Rehabilitation Medicine, 37*, 23-31.

MacElroy, B. (1999). Comparing seven forms of online surveying. *Quirk's Marketing Research Review*. Retrieved December 2, 2002, from http://www.quirks.com/articles/article.asp?arg_ArticleId=510

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics, 3*(4), 201-230.

Norman, K. L. (1991). *The psychology of menu selection: Designing cognitive control at the human/computer interface*. Norwood, NJ: Ablex Publishing Corporation.

Pettit, F. (2002). A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behavior Research Methods, Instruments, & Computers, 34*(1), 50-54.

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

Rest, J. R., Narvaez, D., Thoma S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology, 91*(4), 644-659.

Rest, J., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology, 91*, 644-659.

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*(1), 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*(3), 498-507.

RUMM. (2000). RUMM Laboratory Pty Ltd.

RUMM 2020. (2004). RUMM Laboratory Pty Ltd.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105.

Thoma, S. J. (2002). An overview of the Minnesota approach to research in moral development. *Journal of Moral Development, 31*(3), 225-245.

Yun, G. W., & Trumbo, C. W. (2000). Comparative response to a survey executed by post, e-mail, & web form. *Journal of Computer-Mediated Communication, 6*(1).

Author Note:

This project was partially supported by a faculty development grant from Tarleton State University in 2005. We are grateful to Jeffery L. Kirk, and our anonymous reviewers for their insightful contributions to this paper.