Adaptation and Validation of a Test of Ethical Sensitivity in Teaching[1]

Bruce Maxwell, Professor[2]
University of Quebec Trois-Rivières

Helen Boon, Associate Professor[3]
James Cook University Townsville

Nicolas Tanchuk, Assistant Professor[4]
Iowa State University

Bryan Rauwerda[5]
James Cook University Townsville

[1] This research received financial support from the Social Science and Humanities Research Council of Canada (grant number: 430-2016-01208).

[2] Bruce Maxwell is Professor of Education at the University of Quebec Trois-Rivières, 3351 boul des Forges, Trois-Rivières, Quebec, Canada G8Z 4M3; email: <bruce.maxwell@uqtr.ca>. His research focusses on ethics and law for educators, Quebec's statutory ethics and world religions curriculum and ethical issues in education.

[3] Helen Boon is Associate Professor in the College of Arts, Society and Education at James Cook University, Townswille, University Drive, Townsville QLD 4810, Australia; email <helen.boon@jcu.edu.au>. Her teaching and research interests focus on educational psychology, special needs, climate change and behaviour management. She also has expertise in quantitative research methods, including statistical modelling and Rasch modelling.

[4] Nicolas Tanchuk is Assistant Professor in the School of Education at Iowa State University, Lagomarcino 901 Stange Rd, Ames, IA 50011, USA; e-mail <ntanchuk@iastate.edu>. His research interests focus on the intersection of political philosophy, normative ethics, and professional ethics in educational decision-making.

[5] Bryan Rauwerda studies education at the College of Arts, Society and Education at James Cook University, Townswille, 373 Flinders St, Townsville QLD 4810, Australia; email <helen.boon@jcu.edu.au>.

This article documents the adaptation, piloting and validation of a measure of teachers' ethical sensitivity. To create the test, we modified a measure from dentistry drawing on literature in teacher professional ethics and drew on the expertise of professional ethics scholars and practitioners. Based on the results of Rasch analysis combined with traditional approaches to psychometric validation, the instrument was found to be a valid and reliable means of discerning levels of ethical sensitivity within the group of participants. However, participants' lack of ethical sensitivity as revealed by overall low scores, we contend, lends credence to concerns that teacher education programs may not be adequately preparing future teachers to meet expectations in connection with the ethical dimensions of the profession. In conclusion, we call for further research on pre- and in-service teachers' capacity to perceive, reason about and react appropriately to ethical situations encountered at work.

Over several decades, research on education students' ethical development has consistently found that pre-service teaching students obtain lower scores on standardized tests of moral reasoning than their peers enrolled in other programs of study (Bloom, 1976; Chang, 1994; Cummings, Dyas, Maddux, & Kochman, 2001; Derryberry, Snyder, & Wilson, 2006; Greer, Searby, & Thoma, 2015; McNeel, 1994; O'Flaherty & Gleeson, 2017; Yeazell & Johnson, 1988). Furthermore, cohort studies of undergraduate students indicate that the moral judgment development of education students does not significantly improve over the course of their programs of study, a trend that runs counter to the typical developmental trajectory of young adults (Bakken & Ellsworth, 1990; Boom & Molenaar, 1989; Cummings et al. 2001; McNeel, 1994; Rest, Narvaez, Bebeau & Thoma, 1999; Ünal, 2011; Yeazell & Johnson, 1988). Beyond the fruits of this research program on educators' moral reasoning development, however, little is known about teachers' ethical development and, in particular, how ethical competency can be shaped by educational experiences during professional formation (Campbell, 2008; Maxwell & Schwimmer, 2016a). Outcome research on how ethics instruction affects such things as skills in reasoned

reflection, the understanding of and ability to apply ethical concepts, and awareness of ethical issues and conflicts that arise in practice requires reliable and valid assessment tools. Yet apart from the Defining Issues Test—the measure used extensively in previous research to assess the development of teachers' ethical reasoning ability—such tools are lacking. Drawing on James Rest's Four-Component Model of Moral Functioning as a theoretical framework (see Rest, 1983; Bebeau, Rest, & Narvaez, 1999), the objective of the research reported in this article was to broaden our arsenal of measures of teachers' ethical development by creating and validating a research instrument for assessing the ethical sensitivity of educational professionals.

Previous Ethical Sensitivity Tests and Instrument Design

While as many as 19 different instruments of ethical sensitivity have been developed, none of them is specific to assessing ethical sensitivity in teaching, few have been extensively validated, and even fewer elicit the construct of ethical sensitivity as it is conceptualized in the FCMM (for a review see You, Maeda, & Bebeau, 2011). Existing measures of ethical sensitivity in teaching tend to take one of two forms: recognition tests or self-report tests. Recognition tests ask the test taker to identify possible ethical issues in a scenario or particular professional responses that a situation may call for from a list of pre-set items. Fedeles' (2004) Teachers' Concerns Questionnaire is an example of a recognition test of ethical sensitivity. Self-report tests measures call upon an individual to appraise their own socio-moral abilities or competencies associated with the concept of ethical sensitivity. Tirri and Nokelainen's (2007, 2011) Ethical Sensitivity Scale Questionnaire is an example of a self-report test of ethical sensitivity in teaching.

Instruments of ethical sensitivity that employ such tick-a-box methods are quick and convenient for researchers to use. However, as You, Maeda, & Bebeau (2011) point out, the ability to pick out from a list the ethical issues that might be at stake in a vignette is a poor proxy for the ability to recognize and name the ethical issues that an ethical situation presents. Real-life situations do not provide such prompts. As for self-report measures, validation studies of such measures (see Gholami & Tirri, 2012; Gholami, Kuusisto & Tirri, 2015; Kuusisto, Tirri & Rissanen, 2012, Tirri & Nokelainen, 2011) do not make them immune from the usual objections about the limits of self-report measures. Asking individuals to provide self-perceptions of ethical sensitivity may be inaccurate and is vulnerable to social desirability, deception, and impression-formation effects.

Hence, the key to designing a test intended to assess spontaneous or intuitive ethical reactions to real-life situations is to avoid the use of tasks that rely on prior interpretations—as in the case of, for instance, a multiple-choice test or ranking of question listing several ethical principles or values. It is for this reason that the existing measures of ethical sensitivity generally considered to elicit the construct of ethical sensitivity well center on so-called "unstructured" ethical problems (Clarkeburn, 2002; Sadler, 2004; You & Bebeau, 2005). An unstructured ethical problem is a relatively short depiction of an interaction, often in the form of a dialogue, in which ethical issues are at stake. The narrative gives multiple situational cues to the ethical issues at stake but does not refer explicitly to ethical concepts such as "fairness," "responsibility," "caring," "wellbeing," and "autonomy."

Two previously developed ethical sensitivity assessments that do elicit the concept of ethical sensitivity well are the Dental Ethical Sensitivity Test (Bebeau, Rest & Yamoor,

1985) and the Racial Ethical Sensitivity Test (Brabeck et al., 2000). We used these instruments as models in the design of the Test of Ethical Sensitivity in Teaching (TEST). Accordingly, the TEST is a situational judgment test that presents the test taker with unstructured ethical problems and, in response to a series of probe questions, participants must themselves identify the ethical issues at stake. As well, to maximize the match between the test's medium and the targeted construct, we followed Brabeck et al. (2000) in using video as a medium for presenting the scenarios. As these authors point out, because ethical sensitivity implies an awareness of verbal and nonverbal situational cues, video is a better stimulus for assessing ethical sensitivity.

Theoretical Framework

The Four-Component Model of Morality

This research adopts James Rest's Four-Component Model of Morality (FCMM) as a conceptual framework (see Bebeau, 2014; Bebeau & Monson, 2008). The FCMM combines various theoretical perspectives on moral functioning into a single model of moral functioning (Bebeau, Rest & Narvaez, 1999). In terms of the model's origins, Rest argued that the various theories of moral functioning vying for dominance in the field of moral psychology during the 1980s—the cognitive-developmental approach, the psychoanalytic approach, the empathy-based approach and the socialisation approach—made unwarranted claims to comprehensiveness (Rest, 1983). In his alternative view, each theory was better conceived as highlighting just one of several aspects of moral functioning. These aspects became the basic constructs of his multi-component model (see Table 1). Moral judgment is the capacity to identify morally right or preferable action choices on the basis of considered reflection (component 2) whereas moral motivation is

synonymous with moral integrity or moral responsibility—that is to say, the prioritization of moral values over other values and action incentives (component 3). If moral character corresponds to questions surrounding the determination to pursue moral goals and overcome impediments to the execution of moral acts (component 4), the moral sensitivity component embraces the perception of situations as presenting a moral problem and imagining and predicting the effects of action alternatives on the welfare of potentially affected parties (component 1).

Much as Rest intended it, the FCMM continues to have taxonomic importance, loosely delineating four branches of moral psychology as a field of empirical research and four corresponding areas of moral-educational intervention (Rest, Narvaez, Thoma & Bebeau, 2000).

Table 1

James Rest's Four-component Model of Moral Psychology

Component	Description
1. Moral sensitivity	Perceiving a situation as presenting a moral problem, imagining and predicting the effects of action alternatives on others' welfare
2. Moral judgement	Identifying morally right or preferable actions on the basis of considered reflection
3. Moral motivation	Moral integrity or moral responsibility, consistency between moral judgement and moral action
4. Moral character	Personological factors that affect the agent's determination to execute moral actions and pursue goals, strength of will to resist impediments like fatigue, distractions and setbacks

The Construct of Moral or Ethical Sensitivity

In the FCMM, moral sensitivity describes the capacity to generate an initial interpretation of what is at stake when faced with a moral situation (Rest, 1983). More

specifically, moral sensitivity involves perception in connection with four morally salient aspects of social situations that generate a sense of uncertainty about what to do from a moral standpoint: the set of possible action alternatives, the probable consequences of action alternatives on different parties affected by them, the rights and responsibilities of the various actors involved in the situation, and mitigating or aggravating circumstantial factors (Rest, 1986). Understood this way, moral sensitivity relies on perspective taking—consciousness of how people will be affected by each course of action and how the different players in a situation would regard the effects on their welfare and interests.

The Restian conceptualization of moral sensitivity employed in this study differs from more recent conceptualizations in virtue of its agnosticism about whether a commitment to certain moral principles and values, caring in particular, is integral to the notion of moral sensitivity. Of note in this connection is the model of ethical sensitivity that Tirri and colleagues' adopt in their research program on ethical sensitivity in teaching (see Tirri, 2019 for a summary). Following Narvaez and Endicott (2009), Tirri and colleagues' work complements the Restian conceptualization of moral sensitivity with the skills of "caring and connecting to others," "preventing social bias" and "working with interpersonal and group differences." In our study, we elected to adhere to a more classical interpretation of the FCMM which, in our interpretation, associates caring and working to achieve moral goals like preventing bias and discrimination with moral judgement (component 2) and moral motivation (component 3). On this interpretation, moral sensitivity is conceptually closer to cognitive empathy than affective empathy. Broadly construed, cognitive empathy refers to the ability to perceive other's mental states (beliefs, feelings, desires, etc.) whereas affective empathy implies experiencing emotional reactions

that are more appropriate to another person's situation than to one's own (Hoffman, 2000). Finally, because the Restian conceptualization of moral sensitivity implies consciousness of the norms and rules that apply, it also embraces background knowledge of how certain circumstances generate social expectations in the form of moral obligations (You, Maeda & Bebeau, 2011).

This deontological dimension of ethical sensitivity led Bebeau (2002) to coin the term "ethical sensitivity" to single out a particular sub-category of moral sensitivity. According to Bebeau's (2002) definition, ethical sensitivity is moral sensitivity exercised in a specific professional context. Its use signals that what is at issue are distinctive professional expectations. These expectations are often derived from codes and law that govern professional conduct but may also link to informal norms of practice. The present study employs "ethical sensitivity" in this sense.

Methods

Scenario Development

To develop the scenarios, we followed the approach adopted by Bebeau, Rest and Yamoor (1985) in their initial elaboration of the DEST by creating a set of scripts based on professionals' reports of frequently occurring ethical problems in the workplace. In the case of the DEST, the scripts drew on the results of a preliminary stage of research in which extensive interview data from practicing dentists was collected and analyzed (see Bebeau, Reifel, & Speidel, 1982). In the case of our study, research on teachers' perceptions of ethical issues arising in practice (i.e., Barrett, Casey, Visser, & Headley, 2012) and codes of teacher ethics (Maxwell & Schwimmer, 2016b) allowed us to leapfrog this stage.

The previous research on recurrent ethical issues in teaching informed the thematic content of a set of scenarios representing a spectrum of the most common and ethically difficult problems encountered by practicing teachers (i.e., Barrett, Casey, Visser, & Headley, 2012; Maxwell & Schwimmer, 2016b). See Table 2. Based on the results of this research, we began by drafting four short scenario summaries and subsequently rewrote them in the form of dialogic scripts. The dominant ethical themes in these initial versions of the scenarios were the duty to report suspected cases of abuse or neglect (Parental Meeting), teachers' right to a private life (Religious Symbols), confidentiality and respect for students (Faculty Lounge), and professional autonomy (Reading to Grade Level). The ethical problems represented in each scenario also dovetail with the moral dilemma and conflict categories established in Tirri and Husu's research on teachers' moral and ethical dilemmas in schools (see Husu & Tirri, 2001; Tirri, 1999; Tirri & Husu, 2002). Their research revealed that the moral dilemmas that teachers commonly encounter in the course of their work can be divided into four categories: matters related to teachers' work (managing pupil behavior, confidentiality, unprofessionalism, etc.), pupil's attitudes towards school, the rights of minority groups, and rule enforcement (Tirri, 1999). They also found that conflicts over ethical issues in schools divide naturally into three types: conflicts between teachers and parents about what is in the child's best interest, between teachers and colleagues about the proper exercise of authority, and between individual teachers and the ethical culture of the school community (Husu & Tirri, 2001; Tirri & Husu, 2002). Verbal descriptions of the two scenarios retained for the final version of the TEST appear below in Annex 1 (see the section below Modifications based on the pilot phase).

Table 2

Research-based thematic content of the scenarios

	FACULTY LOUNGE	PARENTAL MEETING	READING TO GRADE LEVEL	RELIGIOUS SYMBOLS
Frequently occurring unethical behavior (Barret et al., 2013)	Gossips to other teachers about a student	Fails to report a colleague's ethical behavior	A teacher does not follow curriculum guidelines	Behaves in an unprofessional way while outside of work
Recurrent content in codes of professional conduct (Maxwell & Schwimmer, 2016b)	Treat students fairly, respectfully and avoid discrimination	Follow legal protocol in reporting suspected abuse or neglect	Observe respect for authority and workplace hierarchy	Manage criticisms and complaints respectfully and through proper channels
Relational conflict category (Husu & Tirri, 2001)	Cultural conflict	Conflict between teachers and parents	Conflict between teachers and colleagues	Cultural conflict
Common moral dilemmas (Tirri, 1999)	Morality of pupil's behavior regarding school and work	Matters related to teachers' work	Matters related to teachers' work	Rights of minority groups

Once drafted, the scenario dialogues were vetted during a series of discussion meetings with seven teacher-partners and four academics with expertise in the area of professional ethics in three countries (the United States, Canada and the Netherlands). The meetings took place by videoconference. Professional relevance and realism—both in terms of the language used and the situations depicted—were checked for and modifications were made to the scenario scripts in accordance with feedback received during the meetings. Next, four-minute long animated video versions of the refined scenarios were produced and a web-based test portal was set up.

Probe Questions

Concurrently with the development of the test scenarios, a set of probe questions was written up based on the questions used in the DEST. The probe questions were crafted to prompt respondents to articulate and justify their intuitive reactions to the scenarios and, in particular, draw the test-taker's attention to the four key aspects of the ethical sensitivity construct (see Table 3). As mentioned above, the four aspects are action alternatives, action consequences, rights and responsibilities, and circumstantial factors. The wording of an initial version of the probe question was also verified by the international team of teacher-partners for clarity and meaningfulness and was modified in light of their comments.

Table 3

TEST Probe Questions

1. Imagine that you are the person identified at the end of the video. Taking on that person's role, what would you say? Respond as if you were speaking directly to other people in the video.
2. Why would you respond in the way you have indicated? What are your reasons for responding this way?
3. How do you think the other people in the situation would understand and react to what you said? Say why you think they would react that way.
4. What are the issues in this situation? In other words, whose well-being, interests or feelings are at stake and what values seem to come into play?
5. What are different peoples' responsibilities in this situation? In other words, who has obligations to which people, what are those obligations and what rights are at stake?

Scoring System

The scoring system also benefitted from input from the project's teacher-partners. The first step in developing the scoring system was to create lists of ethically salient features of each scenario under three separate headings corresponding with the dimensions of the ethical sensitivity construct: rights and responsibilities, circumstantial factors, and action alternatives and consequences. Devising the scoring items was a highly iterative, collaborative and interpretive process that aimed at consensus among the members of the

research team on the nine most ethically salient features of each scenario. Once consensus was reached, a scoring checklist was created for use by the rater to help determine whether written response sets displayed evidence of recognition of the scenario's nine ethically salient features. The final version of the response items for the scenarios titled Faculty Lounge and Parental Meeting, with their corresponding item labels used for the purpose of data analysis, is given in Table 4. These scenarios were the ones used in the final version of the TEST (see the section below Modifications based on the pilot phase).

Table 4

Item Labels and Descriptions for the Scenarios

	FACULTY LOUNGE	
Rights and responsibilities (item subscale)		
1	FL_RR_1	Staff members have a responsibility to work together to find a solution to the student's disruptive behavior
2	V4_A	Teachers should not engage in racial/ethnic/social/economic stereotyping
3	V5_A	Teachers should discuss their difficulties and frustrations with pupils and their families in a respectful way
4	V6_A	Teachers should exercise discretion in the exchange of private or personal information about pupils and their families.
Circumstantial features (item subscale)		
1	FL_CF_2	The students' negative behavior is likely due to a very difficult situation at home (e.g., poverty or insufficient parental supervision) or other causal factors beyond his control (e.g., learning difficulties, ADHD)
2	V8_A	Since the pupil's teacher has just returned to work after maternity leave and is tired, her judgement about the situation may be impaired
3	V9_A	Faculty lounges are not necessarily private spaces
Action impacts (item subscale)		
1	FL_AI_3	The teachers' inability to manage the pupils' disruptive behavior make it difficult for him to benefit from educational opportunities
2	V11_A	If the bystanding teacher objects to or openly criticizes his colleagues' behavior, this will have a negative impact on his relationship with colleagues
	PARENTAL MEETING	
Rights and responsibilities (item subscale)		

1	PM_RR_1	The teachers have a responsibility to act in the best interest of the pupil and his family by bringing the dangerous/harmful situation to the attention of school staff/child protection
2	V19_A	The teachers have a duty to respect the mother's decision/her autonomy in the situation
3	V20_A	Jacob's teacher has a responsibility to keep the promise she made to his mother to keep what she told her about the situation at home confidential
4	V21_A	The bystanding teacher promised that what her colleague told her would remain confidential
Circumstantial features (item subscale)		
1	PM_CF_2	This situation at home is having a very negative impact on the pupil's personal well-being and/or his ability to learn and participate at school
2	V23_A	The pupil and his mother are living in a dangerous home situation which could potentially degenerate
3	V24_A	Teachers and school staff must have sufficient evidence of abuse at home before reporting a suspected situation
Action impacts (item subscale)		
1	PM_AI_3	If the pupil's teacher does not keep her promise to confidentiality, this will likely have a negative impact on her relationship with the mother (loss of trust, feeling of betrayal) ;
2	V26_A	Reporting the situation at home could have a negative impact on the family (removal of the child, increased threat of violence from the father)

The rater's task was to read each response set carefully and, for each ethically salient feature indicated on the scoring checklist, assign one point if the response set displayed any reasonable evidence of recognition and no point if the rater found no such evidence. For each of the scenarios, there were four items under the "rights and responsibilities" subscale, three items under the "circumstantial features" subscale and two items under the "action alternatives and consequences" subscale. Hence, response sets could obtain a maximum of nine points for each scenario.

Test Format

With the scoring system established, the test was then set up for distance use on an online survey platform. In addition to the videos depicting the scenarios, the probe questions and an information and consent form, the questionnaire contained a set of

demographic questions pertaining to such personal variables as career stage and orientation, educational background, religiosity, and employment status. It also included a set of four standard logical problems to probe participants' reasoning ability.

Results

Pilot Phase

The aims of the pilot phase were to check for inter-rater reliability, obtain evidence of construct validity, and refine the scoring system and probe questions. Snowball sampling, a non-probabilistic approach to recruitment, yielded the targeted number of participants in the two participant categories. For this phase of the study, the test platform randomly assigned two of the four original scenarios. Taking into account incomplete response sets, the pilot phase yielded 18 to 21 complete written response sets per scenario.

Participant information. The pilot respondents, 74 early career teachers and education students at various program stages, were located in four OECD countries, 40% from Australia, 40% from Canada, and 10% from the Netherlands and the United States respectively. Consisting primarily of relatively young adults, 55% of the group of participants were 25 to 40 years of age, 25% were under 25 and 20% were over 40, with women making up a majority of the sample (70%). About a third of participants were students enrolled in a program leading to teacher certification. The rest were practicing teachers. Of the practicing teacher participant, 35% taught at the primary level, 60% at the middle- or high-school level and 5% worked in special needs education. **Scoring and inter-rater reliability.** Three members of the research team independently coded and scored each complete response set for all four scenarios. Ordinal results were entered

manually into SPSS for analysis. As mentioned above (see Scoring System), for each of the scenarios' nine items, one point was assigned for evidence of recognition of a relevant factor. A score of zero was recorded if there was no evidence of recognition in the response set but a response was offered to distinguish from missing values which were coded "99".

To check for inter-rater reliability, intra-class correlation coefficients (ICC) were calculated to assess degree of consistency of raters. Intra-item correlations were also calculated to assess degree of absolute agreement between raters' scores, or the degree to which two or more raters achieve identical results under similar assessment conditions.

The four scenarios provided varying degrees of absolute agreement between the raters showing whether the raters' scores are interchangeable. For each of the four scenarios, the reliability analyses generated ICC values for consistency and absolute agreement that were indicative of good or excellent reliability (see Table 5). The highest absolute agreement was obtained for Reading to Grade Level, indicating that this scenario generated the least ambiguity between raters, the highest consistency of ratings, and their scores for this scenario were most interchangeable. The only scenario that fell below the excellence threshold of 0.90 was Parental Meeting.

Table 5

Inter-Rater Reliability Analysis

Scenario	N	Consistency ICC (intra-class correlations based on means, 95% CI)	Absolute agreement ICC (inter-item correlations based on means, 95% CI)
Religious symbols	19	.935[95% CI :.861 - .973]	.907 [95% CI: .752- .965]
Faculty lounge	21	.893 [95% CI:.790-.953]	.818 [95% CI : .474 - .931]
Parental meeting	18	.776 [95% CI:.508 - .910]	.785 [95% CI : .522 - .914]
Reading to grade level	20	.938 [95% CI:. 870 - .974]	.925 [95% CI:.830- 0.969]

Rasch analysis and item validation for the pilot phase. Fit statistics obtained through Rasch analyses were used to help us to select the best scenario based on the suitability of the scenario items to measure the underlying construct, ethical sensitivity. Annex 2 displays the fit statistics, person reliability and item reliability (Saidfudin et al., 2010). Any reliability value which is close to 1 is considered consistent internally (Oon et al., 2017). This indicates that the items are supposedly measuring the trait as required, while a high separation index of 2.12, for example, exceeds the cut-off point of 2.0 suggested by Fisher (2007). A good item reliability means that the sample was big enough to precisely locate the items on the latent variable. Item reliability depends on the item difficulty variance such that a wide difficulty range yields high item reliability. Person reliability depends chiefly on the sample's ability variance such that a wider ability range will result in a higher person reliability. The person's separation index refers to the spread of all the respondents along the continuum measured by the construct's items. Bearing in mind that the pilot samples were perforce very small we selected the scenarios for the validation phase which achieved the highest item reliabilities, namely Faculty Lounge and Parental Meeting.

The three of the items in the TEST that did not fit the Rasch model all fell into the item category of rights and responsibilities. According to the model, respondents found two of them too easy to identify—PM_RR_1, which relates to the duty to report, and V20_A, which relates to the duty to keep promises—and one was too difficult—namely, V6_A linked to the duty to exercise discretion in the exchange of personal information about pupils. Item descriptions can be found above in Table 4. Despite these results, we felt that we could not eliminate these items from the scoring guide and respect the

conceptual integrity of the scenarios. In the case of Parental Meeting, the central ethical problem faced by the protagonist is structured by a tension between precisely these two duties, the duty to keep promises and the duty to report. Recognizing these two items is essential to recognizing that the scenario presents an ethical problem at all. For a similar reason, we elected not to modify the Faculty Lounge scoring guide in light the item validation results. As the teacher partners who helped construct and validate the scoring guide emphasized (see Scoring System), recognizing a teacher's duty to report is not the only crucial matter to understanding the issues that are at stake in the scenario. The admonition to exercise discretion in the exchange of private information about one's pupils is also one the most frequently recurring items in codes of professional conduct for teachers (Maxwell & Schwimmer, 2016b).

Modifications based on the pilot phase. The first phase of the study revealed three principal limitations of the pilot version of the TEST. One was that the participants may have found the demographic section of the questionnaire too time-consuming as demonstrated by the significant number of incomplete responses. Missing responses were 5% for Religious Symbols, 45% for Faculty Lounge, 60% for Parental Meeting and 24% for Reading to Grade Level. Also, the choice to randomly assign participants two of the four scenarios complicated the scoring process and raised questions about comparability between participants' sensitivity scores. Finally, the approach to scoring adopted for the pilot phase, which involved recording scores outside the survey platform then entering them into SPSS by hand (see the above section Scoring and inter-rater reliability), made scoring laborious.

In response to these limitations, the set of demographics on the questionnaire was pared down and the lack of integration between the online test platform and SPSS was corrected. As for the choice of scenarios, for the validation phase, participants were presented with two fixed scenarios, Faculty Lounge and Parental Meeting. This choice was based on the fact that, of the four original scenarios, Rasch analysis revealed that Faculty Lounge and Parental Meeting were the best instruments to measure the underlying construct, ethical sensitivity, as mentioned above.

Validation Phase

Recruitment. With the aim of obtaining evidence of predictive validity in mind, we sought a cross-sectional sample of educators in seven participant categories: beginning education students (131), finishing education students (68), early-career teachers (47), mid- to late-career teachers (54), educational administrators (8), teacher educators (23) and exemplary educators (18). Five participants did not indicate their career stage. We relied on self-identification for the purposes of categorization. The rationale for selecting these participant groups is presented below under Predictive Validity.

Obtaining a sufficient number of respondents in each of the participant categories required a multi-pronged approach to recruitment. For all participant groups except that of exemplary educators, the recruitment strategy began with snowball sampling. A non-probabilistic approach, snowball sampling yielded the targeted participant numbers for the teacher educator category and the two student categories. Members of the research team commissioned the help of colleagues working in teacher education to have their current students complete the TEST or to arrange for a participation invitation to be sent to all

students in their academic unit meeting the inclusion criteria. Attempts at snowball sampling to reach the three categories of in-service educators—early-career teachers, mid- to late-career teachers and educational administrators—proved difficult. To achieve the projected participant numbers we resorted to placing recruitment advertisements targeting in-service teachers on a social media platform. Despite these efforts, we were unable to recruit sufficient numbers of educational administrators. To reach educators falling into the category of exemplary educators, we sent invitations to all winners of recognized teaching awards in the United States (the Council of Chief State School Officers' National Teacher of the Year Program), Canada (Prime Minister's Award for Teaching Excellence) and the six Australian states whose Departments of Education or Colleges of Teachers offer teaching awards: New South Wales, Queensland, South Australia, Tasmania, Victoria, and Western Australia.

Sample characteristics. Of the 354 educators who participated in the study, 39.3% were from Australia, 44.9% were from Canada, 0.6% from Singapore and 15% were from the United States of America. Approximately three quarters of the participants were women (76.8%) and a strong majority of respondents were relatively young educators. Forty-five and a half percent were under 27 years of age, 31.3% were between 27 and 41 and 23.2% were over 42.

Coding and scoring of participant responses. Following the two-step scoring system described above (see Scoring System), one member of the research team read through the participants' written answers to the probe questions with the aim of identifying reasonable textual evidence of recognition of the ethical issues present in the scenarios. For each of the nine items under each of the three sub-scales, coding this qualitative data using

the pre-prepared scoring checklists, the rater assigned as score of 1 if such evidence was found and a score of zero if the response set displayed no evidence of recognition.

Demographic variables and sensitivity score associations. In order to check whether any of the demographic characteristics of the sample predicted ethical sensitivity, stepwise regression analyses were performed using the variables Gender, Reasoning Score, Religiosity, Ethics Courses Studied, Teaching Qualifications, Career Stage, Excellence Award Group, Country and Teaching Context (Elementary, Special Education or Secondary Education). Results of the analyses (see Tables 6 and 7) show that the only variables that significantly predict ethical sensitivity scores are Gender (coded 1 for females and 2 for males), Reasoning Score, and being situated in Canada (Country, coded as separate dummy variables for Canada, Australia and USA, with Singapore as the reference variable). The results of the Rasch measure of ethical sensitivity using country code as a dummy variable are presented in Annex 3. The results show that although these three variables—Gender, Reasoning Score and Country—are significant predictors of ethical sensitivity, only 6% of the variance in ethical sensitivity is predicted by model 3. Also note that the stepwise regressions comparing the other Country coded dummy variables showed they were not significant predictors of ethical sensitivity (Table 7). The results show that Canadian participants demonstrated greater ethical sensitivity than respondents from other countries.

Table 6

Stepwise Regression Analysis for the Interval Measure of Ethical Sensitivity

Model	*Step and predictor variable*	*B*	*S.E.*	*Beta (β)*	R^2	ΔR^2	*p*
1	(Constant)	-.809	.050				0.000

	CANADA (v Reference Singapore)	.238	.072	.174			
					0.30[a]	0.30	0.001
2	(Constant)	-.959	.074				0.000
	CANADA (v Reference Singapore)	.239	.071	.175			0.001
	REASONING QUESTION TOTAL	.080	.029	.142			
					.051[b]	.020	0.007
3	(Constant)	-.767	.122				0.000
	CANADA(v Reference Singapore)	.276	.073	.202			0.000
	REASONING QUESTION TOTAL	.081	.029	.144			0.006
	GENDER	-.177	.089	-.106	.061[c]	0.011	0.048

a. Predictors: (Constant), CANADA (v Reference Singapore)
b. Predictors: (Constant), CANADA (v Reference Singapore), REASONING QUESTION TOTAL
c. Predictors: (Constant), CANADA (v Reference Singapore), REASONING QUESTION TOTAL, GENDER
NB. B, unstandardized coefficient; S.E. Standard Error; β, standardized coefficient; R^2 variance explained by the model; ΔR^2 change in variance explained at each step of the regression model; *p* , probability.

Table 7

Excluded Variables in Stepwise Regression Analyses for Ethical Sensitivity Associations

Excluded Variables[a]						
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	REASONING QUESTION TOTAL	.142[b]	2.727	.007	.144	1.000
	AUSTRALIA (v Reference Singapore)	.064[b]	.859	.391	.046	.498
	USA(v Reference Singapore)	-.051[b]	-.871	.385	-.046	.817
	GENDER	-.104[b]	-1.923	.055	-.102	.936

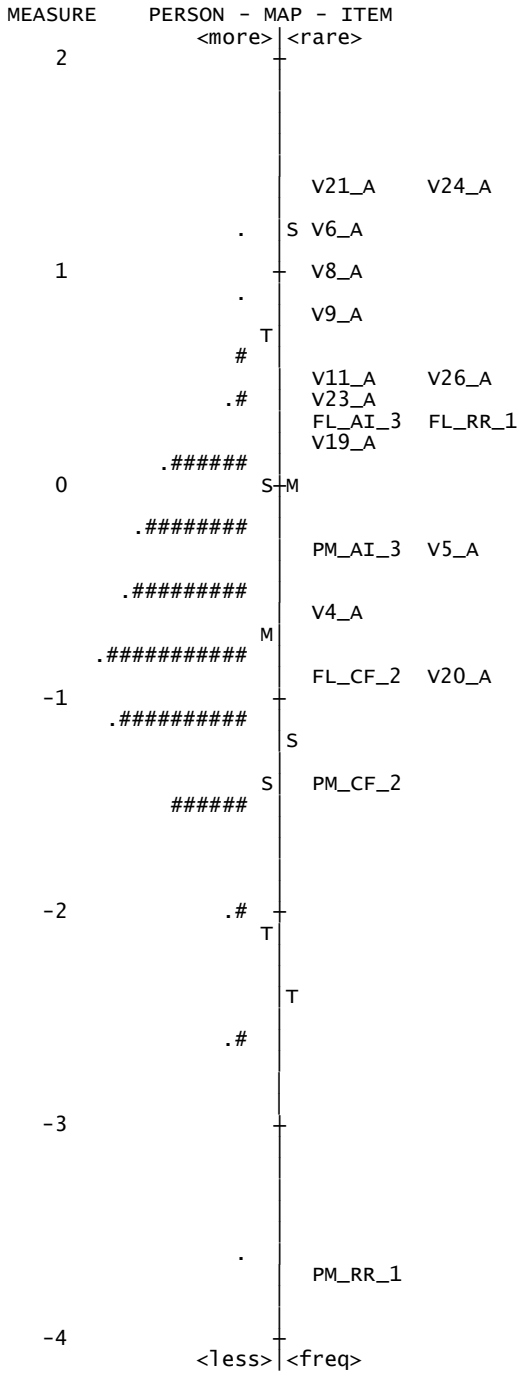
2	AUSTRALIA(v Reference Singapore)	.046[c]	.618	.537	.033	.494
	USA(v Reference Singapore)	-.035[c]	-.604	.546	-.032	.809
	GENDER	-.106[c]	-1.982	.048	-.105	.935
3	AUSTRALIA(v Reference Singapore)	.032[d]	.431	.667	.023	.489
	USA(v Reference Singapore)	-.024[d]	-.407	.684	-.022	.800

a. Dependent Variable: Interval Measure of Ethical Sensitivity
b. Predictors in the Model: (Constant), CANADA(v Reference Singapore)
c. Predictors in the Model: (Constant), CANADA(v Reference Singapore), REASONING QUESTION TOTAL
d. Predictors in the Model: (Constant), CANADA(v Reference Singapore), REASONING QUESTION TOTAL , GENDER

Rasch analysis. Rasch analysis was used to assess the instrument for uni-dimensionality and person-item fit. The original 18 items form a unidimensional scale presented in Figure 1. The person-item map displays distribution of items on the right of the map and distribution of persons on the left. The top represents the hardest items and participants with most ethical sensitivity. By contrast, the bottom represents the easiest items and participants with least ethical sensitivity. On the person-item map, items are considered ideal when their distribution is sufficient to cover the distribution of persons. The 18 items were analyzed using the Rating Scale Model, specifying that a set of items share the same rating scale structure (Linacre, 2014).

Figure 1

Variable Map of Respondents and Instrument Items with Response Thresholds[*]



* Each "#" IS 6 respondents, each "." is 1 to 5 respondents

The aim of the Rasch analysis was to provide a psychometrically sound instrument and the items which were estimated to fit the Rasch model perform this function well.

Table 8 displays the summary of the model fit statistics, mean measure of the construct ethical sensitivity and separations of the items and respondents for the instrument using the validation sample (N=354). The reliability for person and instrument was examined by reliability and separation index. A reliability value above 0.80 is considered good reliability, while a value between 0.67 and 0.80 is fair, and one less than 0.67 is poor (Fisher 2007). A separation index value greater than 3 is considered good. The instrument reliability is very good. However, the separation of the respondents is low and indicates that the instrument was too difficult for the participants.

Table 8

Summary Model Fit, Mean Measure and Separation

SUMMARY OF 354 MEASURED PERSON (non-extreme)									
	Raw score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	
MEAN	6.5	18.0	-.67	.58	1.00	.06	.97	.04	
S.E.	.1	.0	.04	.00	.01	.04	.02	.04	
Max	13.0	18.0	1.22	1.18	1.71	2.20	2.31	2.71	
Min	1.0	18.0	-3.61	.52	.37	-2.11	.08	-1.35	
Real RMSE	.61	TRUE S.D.	0.33	SEPARATION	0.55	PERSON	RELIABILITY	.23	
Model RMSE	.58	TRUE S.D.	0.37	SEPARATION	0.64	PERSON	RELIABILITY	.29	
S.E. of Person Mean	0.04								
SUMMARY OF 18 MEASURED ITEMS (non-extreme)									
	Raw score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	
MEAN	127.7	354	.00	0.13	1.00	.08	.97	-.13	
S.E.	17.7	.0	.29	0.01	.01	.28	.02	.28	
Max	333	354	1.44	0.23	1.11	3.35	1.14	3.13	
Min	44.0	354	-3.66	0.11	.92	-2.05	.80	-1.56	
Real RMSE	.14	TRUE S.D.	1.18	SEPARATION	8.53	ITEM	RELIABILITY	.99	

Model RMSE	.14	TRUE S.D.	1.18	SEPARATION	8.61	ITEM	RELIABILITY	.99	
S.E. of ITEM Mean	0.29								

UMEAN=.0000 USCALE=1.0000

The results in Table 9 indicate that the point biserial correlations ranged from 0.12 to 0.38, indicating that some items link only moderately to the underlying construct (those with a pt.bis corr of less than 0.3). On the other hand the infit (MNSQ) and outfit (MNSQ) statistics are very good, as they adhere to the recommended range of 0.5-1.5 (Bond and Fox 2007).

Table 9

Item Measure, Misfit Statistics and Point Biserial Correlations for 18 Items

		INFIT		OUTFIT		point biserial correlation
Item[1]	MEASURE	MNSQ	ZSTD (T TEST)	MNSQ	ZSTD (T TEST)	
FL_CF_2	-0.89	1.11	3.35	1.14	3.13	0.14[4]
FL_AI_3	0.34	1.09	1.73	1.14	1.74	0.12[4]
V19_A	0.18	1.06	1.2	1.07	0.96	0.19[4]
V11_A	0.49	1.01	0.26	1.04	0.43	0.22
V24_A	1.44[3]	1.03	0.31	1.02	0.16	0.15[4]
FL_RR_1	0.27	1.01	0.16	1	0.01	0.25
PM_RR_1	-3.66[2]	1.01	0.13	0.8	-0.76	0.22
V21_A	1.38	1.01	0.14	1.01	0.14	0.18[4]
PM_CF_2	-1.36	1.01	0.24	1	0.05	0.29
V23_A	0.37	1.01	0.16	1	0.01	0.25
V5_A	-0.34	0.99	-0.17	0.98	-0.44	0.30
PM_AI_3	-0.27	0.98	-0.57	0.95	-1	0.32
V26_A	0.53	0.98	-0.3	0.93	-0.83	0.29
V4_A	-0.55	0.94	-2.05	0.97	-0.7	0.38
V6_A	1.17	0.97	-0.32	0.85	-1.09	0.28

V20_A	-0.91	0.96	-1.22	0.95	-1.2	0.36
V8_A	1.01	0.94	-0.63	0.81	-1.56	0.33
V9_A	0.8	0.92	-1.04	0.85	-1.41	0.36
MEAN	0	1	0.1	0.97	-0.1	
S.D.	1.18	0.05	1.1	0.09	1.1	

[1] See Table 4 for item descriptions, [2] easiest overall, [3] hardest overall, [4] low contribution to the underlying construct of ethical sensitivity

Predictive validity. To assess the validity of the TEST, one approach we adopted was to seek evidence for predictive validity. Predictive validity describes how well a test predicts an examinee's outcome on another measure that assesses performance in a relevantly similar domain or how well it predicts performance or behavior in other comparable situations (McIntire & Miller, 2005). In short, the predictive validity of a psychometric test depends on the extent to which it predicts what it is supposed to predict.

The reference points for gathering evidence for predictive validity were based on informed guesswork about teachers' ethical development supported by previous findings on professional identity development. We hypothesized that participants would become more ethically sensitive through in-service teacher education as they become increasingly versed in the ethical norms and expectations associated with teacher professionalism and that ethical sensitivity would then decline among early-career teachers. The latter assumption drew on the finding that the early-career period is characterized by "survival mode" in which teachers are intensely focused on lesson preparation and class management and are less attentive to other professional demands like communication with parents and colleagues, holistic pupil well-being and contributing to the profession through service work (Bebeau & Monson, 2012; Mukamurera & Tardif, 2016). We conjectured further that, after the early-career period, ethical sensitivity would gradually increase again as

teachers progress through their careers and exposure to the ethical challenges of teaching accumulates accordingly. Finally, we thought it likely that teachers singled out by their communities and peers as exceptionally committed to the craft of teaching would likely display the highest degree of ethical sensitivity. This assumption was based on repeated assertions in the conceptual literature on teacher professionalism and teacher education that quality teaching and the internalization of ethical norms in teaching are intimately linked (Maxwell & Schwimmer, 2016a). Results obtained by participants based on raw scores of ethical sensitivity by career group are displayed in Table 10.

Table 10

Sensitivity Score by Career Stage

	N	%	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Beginning education student	131	37.0	6.22	1.97	0.17	5.88	6.56	1	11
Finishing education student	68	19.2	6.38	2.31	0.28	5.82	6.94	2	13
Early career teacher < 7 years	47	13.3	6.74	1.97	0.29	6.17	7.32	1	11
Mid to late career teacher >7 years	54	15.3	6.69	2.03	0.28	6.13	7.24	2	12
Exemplary educators	18	5.1	7.33	2.06	0.49	6.31	8.36	4	12
Teacher educator	23	6.5	6.96	1.52	0.32	6.30	7.61	4	9
Educational administrator	8	2.3	5.75	2.66	0.94	3.53	7.97	3	10
MISSING CASES	5	1.4							
Total	354	100	6.49	2.05	0.11	6.27	6.71	1	13

The results of the ANOVA analysis of sensitivity score by career stage did not yield significant differences between the seven participant groups (Table 10). Out of a maximum of 18 points, the mean score for all participant groups was 6.49 (2.05 SD). For exemplary educators, the highest-scoring group, it was 7.33 (2.06 SD).

Variability among participants. Variability across individuals is central to the definition of ethical sensitivity (Bebeau, Rest & Yamoor, 1985) and the very purpose of a test of ethical sensitivity is to assess individuals' levels of ethical sensitivity with a view to propose interventions by way of either professional development or tertiary courses. Though the TEST did not confirm the hypothesis that career stage predicts ethical sensitivity, its use did allow us to observe variability in ethical sensitivity within the sample. Total scores ranged from 2 to 12 out of 18 possible points. The mean was 6.96 ± 2.27 (SD) in a positively skewed distribution.

Discussion and Conclusion

Certain reservations notwithstanding, we believe that the results of this study indicate that the TEST is a valid and reliable measure of educators' ethical sensitivity.

First, we are confident that the scenarios were realistic, accessible and representative of commonly encountered ethical situations and that the scoring system was complete, fair and accurate. The ethical problems were identified based on previous research on recurrent ethical issues in teaching and the details of the scenarios and the scoring criteria were selected and elaborated on in collaboration with an international group of 12 educators representing a range of career stages, teaching areas and levels (see Scenario Development and Scoring System). Two educational administrators were also

part of this group. The successful implementation of this collaborative approach to test development supports the TEST's content validity.

Although the results of our investigation into the instrument's predictive validity disconfirmed our working hypothesis that teachers' ethical sensitivity tends to increase as they gain professional experience, this finding may not be as dire as it appears. The background evidence we used to formulate the hypothesis that career stage has pronounced impact on ethical sensitivity was not strong (see Predictive Validity) and other aspects of the adaptation and validation process indicate promise for the TEST. In addition to its grounding in teachers' professional experiences, as just mentioned, the results of the Rasch analysis in particular show that the items link reasonably well to the underlying construct and the instrument was good as determined by the reliability and separation indices. The TEST would be improved if the five items which had poor loading onto the underlying construct ethical sensitivity (namely FL_AI_3, FL_CF_2, V24_A,V21_A, V19_A) were either reworded or removed from the TEST as they appear to be loading to the component of moral judgment of the FCMM (see Theoretical Framework) rather than moral sensitivity as determined by the point biserial correlations and the wording of those items. As conceptual overlap and interaction between the four components of morality is integral to the FCMM (Rest, 1983), this result is consistent with the measure's theoretical framework.

The TEST may not have differentiated well between educators at different career stages but it did, we saw, observe variability within the sample of participants (see Variability among Participants). Furthermore, our finding that ethical sensitivity levels do not increase with gains in age and education level is consistent the results of a longstanding research program on teachers' moral judgement development. As Cummings, Harlow and

Maddux (2007) report, studies that have examined principled moral reasoning among education students show no improvement from the beginning to the end of their program of professional preparation as measured by the DIT. As noted above, these findings stand in striking contrast with the results of decades of DIT research on college and university students which have found significant increases in moral reasoning scores as students advance in age and education level (Bebeau, 2002; Greer, Searby & Thoma, 2015).

Another result of this study that speaks in favor of the TEST's measurement validity is the finding on gender differences in ethical sensitivity. The notion that women are more morally sensitive than men is, of course, something of a cliché. Taking this received idea as a starting point, You, Maeda and Bebeau (2011) synthesized the conclusions of 19 primary studies on moral sensitivity that included gender as a variable and that met their criteria for statistical sufficiency. The application of meta-analytic techniques resulted in the observation that, on average, female participants score higher on measures of moral sensitivity than male participants. In our study, among the large number of personal variables we collected data on—including religiosity, education level, route to teaching certification, age, and previous college-level coursework—gender was one of the few that correlated significantly with ethical sensitivity. (The others were geographical location and reasoning ability.) We consider this convergence of our findings on gender differences in ethical sensitivity as an indicator of the TEST's concurrent validity.

Finally, in connection with the instrument's practical use value, we would be the first to admit that the TEST has clear disadvantages in comparison with standard paper-and-pencil psychological measures. However, as explained above (see Previous Ethical Sensitivity Tests and Instrument Design), the choice of an open situational judgment format

with prompts for written responses over presenting examinees with lists of items to select from was necessarily to maintain the integrity of the measure's theoretical grounding in the ethical sensitivity construct as is it is understood in the FCMM. Despite this limitation, the scoring system seems relatively easy to master as witnessed by the fact that acceptable levels of interrater reliability were achieved with little training. This feature of the TEST bodes well for its use in educational settings as an instructional aid for exploring the ethical themes central to the scenarios—namely, confidentiality (Faculty Lounge), the duty to report (Parental Meeting), academic freedom (Religious Symbols) and professional autonomy (Reading to Grade Level)—and, more broadly, introducing education students to the notion of ethical sensitivity as a professional competency. Indeed, the scoring system's ease of use suggest that students would be able to score their own and other's responses to the scenarios in class using the checklist.

While our analyses indicate that the TEST is a valid and reliable measure of ethical sensitivity for educators, the fact that participants clearly found the TEST difficult is cause for concern. Specifically, the low overall scores, the fact that the item difficulties are skewed towards the upper end and that some items were recognized only by a very small number of participants raise questions about the TEST's usability with participant groups whose level of ethical sensitivity is expected to be low—beginning education students, most notably. Undoubtedly, the fact that the TEST is a production measure rather than a recognition measure (see Previous Ethical Sensitivity Tests and Instrument Design) contributes significantly to its difficulty. As Gibbs et al. (1992) observe in a discussion of moral functioning measures, one of the practical disadvantages of production measures is that they are more time consuming and complex both for participants to take and for

researchers to score. A classic example is Kohlberg's Moral Judgment Interview (MJI) (Kohlberg, 1981). Furthermore, unlike the MJI, which allows researchers to ask follow up questions thus enabling participants to elaborate on their responses to the interview questions, the TEST relies solely on written responses that participants enter anonymously on an online test platform. In these conditions, some participants may be tempted to respond to the probe questions with short and superficial answers that may not necessarily accurately reflect their thinking.

One way to lower the TEST's difficulty, which we considered but ultimately rejected following the pilot phase, would have been to broaden the coding scheme. At issue, in essence, was a choice between the practical advantages of a more sensitive measure of ethical sensitivity versus maintaining the conceptual integrity of the measure in terms of the degree to which the coding scheme captures the scenarios' inherent ethical features. The purpose of involving teacher-partners in the project was not as much to legitimize the scoring system as it was to ensure the scoring system's qualitative exhaustiveness (see Scoring System). That is to say, in the manner of the Delphi method (Gordon, 1994), the teacher-participants played the role of a panel of experts tasked with reaching consensus on the scenario's observable features in relation to the three dimensions of ethical sensitivity (i.e., rights and responsibilities, ethically relevant circumstantial features, and foreseeable action impacts). A highly iterative process, developing the scoring system development took time and required careful reflection on the various ethical dimensions of the scenarios. We are aware that this is a luxury that the real test-taking conditions do not afford. However, broadening the coding system would have meant reducing the measure's ethical integrity (see Rasch Analysis and Item Validation for the Pilot Phase). Viewed from

this perspective, the TEST's difficulty could be viewed as a strength rather than a weakness in the sense that a very low score accurately reflects participants' lack of sensitivity to the inherent ethical features of the scenarios presented to them.

With that said, in retrospect, the probe questions' wording could have and should be improved in order to elicit higher quality responses. Another explanation for the low scores relates to the possibility that examinees may have misunderstood the task they were expected to perform. The probe questions were designed to guide participants towards thinking about the scenarios in terms of the three dimensions of ethical sensitivity according to the Restian conceptualization (see Probe Questions). Response sets, however, frequently gravitated around one aspect of the scenario that the participant seemed to consider to be particularly ethically salient—most commonly, the duty to report in Parental Meeting and race-based bias in Faculty Lounge. One interpretation of this pattern is that the respondents understood that they were being tested not on their ability to perceive the various aspects of a situation that make it ethically problematic but rather on their ability to correctly identify the most important instance of professional misconduct in each scenario. If this is the case, participants may have found the test "easier" (i.e., obtained higher overall scores) if they had a more accurate understanding that their task was to demonstrate their ability to recognize as many ethically salient dimensions of the scenarios as possible. On this basis, we would recommend that, in future uses of the TEST, the probe questions be reworded to remove the ambiguity about the task that participants are being asked to perform and, in doing so, make clearer reference to the discrete aspects of the underlying ethical sensitivity construct. In the probe questions' present form, question 5,

which prompts participants to think about the agents' ethical responsibilities, serves as a model in this regard.

In conclusion, we believe that the relatively poor performance of participants on the TEST lends credence to widespread concerns that teacher education programs may not be adequately preparing future teachers to meet public and professional expectations with regard to ethical knowledge and conduct (see Cummings, Harlow and Maddux, 2007; Cummings, Maddux & Cladianos, & Richmond, 2010; Cummings, Wiest, Lamitina, & Maddux, 2003; Maxwell & Schwimmer, 2016a; Truscott, 2018). With due caution not to overstate what can be inferred from the application of a single instrument designed to probe one among other components of ethical functioning (ethical sensitivity) in a study that used non-probability sampling methods, the fact that the overwhelming majority of nearly 400 educators from the United States, Canada and Australia seemed to find it difficult to identify even a fraction of the scenarios' ethically relevant aspects, may be cause of concern. At the very least, this finding would suggest a need for further research on pre- and in-service teachers' knowledge of and ability to apply basic professional obligations that recur in codes of ethics and other ethical concepts as well as their capacity to understand and work through the kinds of ethical situations teachers encounter regularly at work. The TEST gives us one of the tools we need to go forward with such research. Others are the recently-validated Teaching Intermediate Concept Measure (Kerr, 2018), which assesses teachers' mastery of basic ethical duties and principles, and generic measures of moral and professional development. Of particular interests in this context are the Defining Issues Test for moral reasoning ability (Thoma 2006) and the Professional Identity Essay, and flexible assessment of professional identity development that has been used and

validated in the context of medicine (Bebeau & Faber-Langendoen, 2014), law (Hamilton, Monson & Organ, 2012), and dentistry (Bebeau & Monson, 2012).

Survey work has shown that teacher education is comparable with other professions in requiring future education professionals to engage in structured teaching and learning of ethics in the form of a discrete course or module (Maxwell, et al., 2016). What we need to know more about now is what kind of instruction in teacher ethics is having the most impact on teachers' ethical development as professionals, particularly as it relates to the broad pedagogical objectives for ethics education around which scholarly and professional opinion in teacher education has coalesced: familiarizing teachers with ethically relevant concepts, helping them understand their professional obligations, promoting the key values of the teaching profession, raising teachers' awareness about teacher professionalism, increasing their sensitivity to ethical issues in context and improving their skills in ethical reasoning (see Maxwell & Schwimmer, 2016b, Maxwell, et al., 2016). The extensive record of professional ethics education in fields outside teaching is a reason to be optimistic. A lack of knowledge of professional ethics as well as profession-specific deficiencies in the four components of ethical functioning can, after all, usually be remediated through the right kind of instruction.

References

Bakken, L., & Ellsworth, R. (1990). Moral development in adulthood: Its relationship to age, sex, and education. *Education Research Quarterly, 14*(2), 2–9.

Barrett, D. E., Casey, J. E., Visser, R. D., & Headley, K. N. (2012). How do teachers make judgments about ethical and unethical behaviors? Toward the development of a code of conduct for teachers. *Teaching and Teacher Education, 28*, 890-898.

Bebeau, M. J. (2002). The defining issues test and the four component model: Contributions to professional education. *Journal of moral education*, *31*(3), 271-295.

Bebeau, M. J. (2014). An evidence-based guide for ethics instruction. *Journal of Microbiology & Biology Education*, December, 124-129.

Bebeau, M. J., & Faber-Langendoen, K. (2014). Remediating lapses in professionalism. In A. Kalet & C. Chou (eds.), *Remediation in medical education* (pp. 103–127). New York: Springer Science.

Bebeau, M. J., & V. E. Monson (2008). Guided by theory, grounded in evidence: A way forward for professional ethics education. In L. Nucci & D. Narvaez (eds.), *Handbook of moral and character education* (pp. 557–582). Hillsdale, NJ: Routledge.

Bebeau, M.J., & Monson, V.E. (2012) Professional identity formation and transformation across the life span. In A. McKee& M. Eraut (Eds.), *Learning trajectories, innovation and identity for professional development: Innovation and change in professional education* (pp. 135–163). Dordrecht: Springer.

Bebeau, M. J., Reifel, N. M., & Speidel, T. M. (1981, January). Measuring the type and frequency of professional dilemmas in dentistry. In *Journal of Dental Research* (Vol. 60, pp. 532-532).

Bebeau, M. J., Rest, J. R., & Narvaez, D. (1999). Beyond the promise: A perspective on research in moral education. *Educational researcher*, *28*(4), 18-26.

Bebeau, M. J., Rest, J. R., & Yamoor, C. M. (1985). Measuring dental students' ethical sensitivity. *Dental Education, 49*(4), 225-235.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: fundamental measurement in the human sciences*. 2nd ed. Mahwah: Lawrence Erlbaum Associates.

Boom, J., & Molenaar, P. (1989). A developmental model of hierarchical stage structure in objective moral judgments. *Developmental Review, 9*(2), 133–145.

Brabeck, M. M., Rogers, L. A., Sirin, S., Henderson, J., Benvenuto, M., & Weaver, M. (2000). Increasing ethical sensitivity to racial and gender intolerance in schools: Development of the racial ethical sensitivity test. *Ethics & Behavior, 10,* 119–137.

Campbell, E. (2008). The ethics of teaching as a moral profession. *Curriculum Inquiry*, *38*(4), 357–385.

Chang, F. Y. (1994). School teachers' moral reasoning. In J. R. Rest & D. Narvaez (Eds.), *Moral judgment development in the professions: Psychology and applied ethics* (pp. 71-83). Hillsdale, NJ: Erlbaum.

Clarkeburn, H. (2002). A test for ethical sensitivity in science. *Journal of Moral Education, 31*(4), 439-453.

Cummings, R., Dyas, L., Maddux, C. D., & Kochman, A. (2001). Principled moral reasoning and behavior of pre-service teacher education students. *American Educational Research Journal*, *38*(1), 143-158.

Cummings, R., Harlow, S., & Maddux, C. D. (2007). Moral reasoning of in-service and pre-service teachers: A review of the research. *Journal of Moral Education*, *36*(1), 67–78.

Cummings, R., Wiest, L. R., Lamitina, D., & Maddux, C. D. (2003). Teacher education curricula and moral reasoning. *Academic Exchange Quarterly*, *7*(1), 163-169.

Derryberry, W. P., Snyder, H., & Wilson, T. (2006). Moral judgment differences in education and liberal arts majors: Cause for concern? *Journal of College and Character, 7*(4), 1–10.

Fedeles, M. (2004). *The teachers' concerns questionnaire: The development and validation of a measure of high school teachers' moral sensitivity* [Unpublished doctoral dissertation]. University of British Columbia.

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21,* 1095.

Gholami. K., Kuusisto, E. & Tirri, K. 2015. Is ethical sensitivity in teaching culturally bound? *Compere: A Journal of Comparative and International Education, 45*(6), 886- 907. DOI: 10.1080/03057925.2014.984588

Gholami, K. & Tirri, K. 2012. The cultural dependence of The Ethical Sensitivity Scale Questionnaire: The case of Iranian Kurdish teachers. *Education Research International 2012*.

Gibbs, J. C., Basinger, K. S., Fuller, D., & Fuller, R. L. (2013). *Moral maturity: Measuring the development of sociomoral reflection*. Routledge.

Gordon, T.J. (1994). The delphi method. *Futures research methodology, 2*(3), 1-30.

Greer, J. L., Searby, L. J., & Thoma, S. J. (2015). Arrested development? Comparing educational leadership students with national norms on moral reasoning. *Educational Administration Quarterly, 51*(4), 511–542.

Hamilton, N. W., Monson, V. E., & Organ, J. M. (2012). Empirical evidence that legal education can foster student professionalism/professional formation to become an effective lawyer. *University of St. Thomas Law Journal*, *10*, 11.

Hoffman, M. L. (2008). Empathy and prosocial behavior. *Handbook of emotions*, *3*, 440-455.

Husu, J. & Tirri, K. 2001. Teachers' ethical choices in sociomoral settings. *Journal of Moral Education, 30* (4), 361-375.

Kerr, S. (2018). *Developing and testing a teaching intermediate concept measure: a preliminary reliability and validity study* (Doctoral dissertation, University of Alabama Libraries).

Kohlberg, L. (1981). The meaning and measurement of moral development. *Heinz Warner Memorial Lecture Series, Vol. 13*. Worcester, MA: Clark University Press.

Kuusisto, E., Tirri, K., & Rissanen, I. 2012. Finnish teachers' ethical sensitivity. *Education Research International 2012.*

Linacre J. (2014). *A user's guide to Winsteps ministep Rasch-model computer programs*. Chicago: Winsteps.

Maxwell, B., & Schwimmer, M. (2016a). Professional ethics education for future teachers: A narrative review of the scholarly writings. *Journal of Moral Education*, *45*(3), 354-371.

Maxwell, B., & Schwimmer, M. (2016b). Seeking the elusive ethical base of teacher professionalism in Canadian codes of ethics. *Teaching and Teacher Education, 59*, 468-480.

Maxwell, B., Tremblay-Laprise, A. A., Filion, M., Boon, H., Daly, C., van den Hoven, M., ... & Walters, S. (2016). A five-country survey on ethics education in preservice teaching programs. *Journal of Teacher Education*, *67*(2), 135-151.

McIntire, S. A., & Miller, L. A. (2000). *Foundations of psychological testing*. McGraw-Hill.

McNeel, S. P. (1994). College teaching and student moral development. In J. R. Rest & D. Narvaez (Eds.), *Moral Development in the professions: Psychology and applied ethics* (pp. 27–49). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mukamurera, J., & Tardif, M. (2016). Épanouissement professionnel: entre développement professionnel, satisfaction au travail et intention de persévérance durant les premières années d'enseignement. *Former les enseignants au XXIe siècle*, *2*, 113-134.

O'Flaherty, J., & Gleeson, J. (2017). Irish student teachers' levels of moral reasoning-context, comparisons, and contributing influences. *Teachers and Teaching, 23*(1), 59–77.

Oon, P. T., Spencer, B., & Kam, C. C. S. (2017). Psychometric quality of a student Oon, P. T., Spencer, B., & Kam, C. C. S. (2017). Psychometric quality of a student evaluation of teaching survey in higher education. *Assessment & Evaluation in Higher Education*, *42*(5), 788-800.

Rest, J. R. (1983). Morality. In P. H. Mussen, J. Flavell, & E. Markman (eds.), *Handbook of child psychology: Vol 3. Cognitive development, 4th ed*. (pp. 556–629). New York: Wiley.

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

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (2000). A neo-Kohlbergian approach to morality research. *Journal of Moral Education*, *29*(4), 381-395.

Sadler, T. D. (2004). Moral sensitivity and its contribution to the resolution of socio-scientific issues. *Journal of Moral Education, 33*(3), 339-358.

Saidfudin, M., Azrilah, A., Rodzo'an, N., Omar, M., Zaharim, A., & Basri, H. (2010). Easier Learning Outcomes Analysis Using Rasch Model in Engineering Education Research. *Latest Trends on Engineering Education,* 442-447.

Thoma, S. (2006). Research on the Defining Issues Test. In M. Killen & J. Smetana (Eds.), *Handbook of moral development* (pp. 49-67). Hillsdale, NJ: Erlbaum.

Tirri, K. (2019). Ethical sensitivity in teaching and teacher education. In M. A. Peters (Ed) Encyclopedia of Teacher Education (Springer Nature). Springer Science+Business Media. https://doi.org/10.1007/978-981-13-1179-6

Tirri, K. 1999. Teachers' perceptions of moral dilemmas at school. *Journal of Moral Education, 28*(1), 31-47.

Tirri, K. & Husu, J. 2002. Care and responsibility in "the best interest of the child": relational voices of ethical dilemmas in teaching. *Teachers and Teaching, 8*(1), 65-80.

Tirri, K. & Nokelainen, P. (2007). Comparison of academically average and gifted students' self-rated ethical sensitivity. *Educational Research and Evaluation, 13*(6), 587-601.

Tirri, K. & Nokelainen, P. (2011). *Measuring multiple intelligences and moral sensitivities in education*. Rotterdam/Taipei: SensePublishers.

Truscott, D. (2018). Teaching professional ethics and law: Blending the professional expectations and reflective practice approaches. In B. Maxwell, N. Tanchuk & C. Scramstad (Eds.), *Professional ethics and law for Canadian teachers* (pp. 1-14). Ottawa, ON: Canadian Association for Teacher Education.

Ünal, E. (2011). Examining the relationship between pre-service teachers' ethical reasoning levels and their academic dishonesty levels: A structural equation modelling approach. *Educational Research and Reviews, 6*(19), 983–992.

Yeazell, M. I., & Johnson, S. F. (1988). Levels of moral judgment of faculty and students in a teacher education program: A micro study of an institution. *Teacher Education Quarterly, 15*(1), 61–70.

You, D., & Bebeau, M. (2005). Moral sensitivity: A review. Conference paper presented at the Annual Meeting of the Association for Moral Education, Cambridge, MA.

You, D., Maeda, Y., & Bebeau, M. J. (2011). Gender differences in moral sensitivity: A meta-analysis. *Ethics & Behavor, 21*(4), 263–282.

Annex 1

Verbal Description of the TEST Scenarios

Faculty Lounge

Sam, a new teacher at the school, is sitting in the staff room with two veteran teachers, Diane and Sophia. Diane, recently back to work after 6 months' maternity leave, is in tears. Dianne is telling her colleague about how badly a particular pupil in her grade 8 class has acted towards her that morning in class. Dianne describes the boy as a "monster," attributing his behavior to his parents, whom she describes as stupid, chronically unemployed and lazy. Trying to console her colleague, Sophia says that the same boy was in one of her classes last year. She brags, "with that kid, I don't mess around. The second he starts getting on my nerves I just kick him out of class." Overhearing the discussion, and seeing how upset Diane is, other staff members start gathering around to listen. Diane and Sophia continue sharing stories about the pupil and all the trouble he has caused in their classes this year. At one point, Sophia mentions that she'd heard that his English teacher last year had "gotten him Ritalin" to try to solve his behavior problems. "Obviously he needs to take something stronger," Diane says, laughing through her tears. Sophia complains that a big part of the problem is that Alex's parents don't care about their son's learning and that this behavior is "typical" of people from "that community". In Dianne's opinion, Alex's behavior is just a way of deflecting attention from his lack of potential. The bell rings to indicate the end of recess, and the teachers gradually leave the room to return to their respective classes, leaving Diane, Sophia and Sam alone in the

staffroom. Sam, who'd been listening up until now, interjects, somewhat cautiously, "I can see you are frustrated but it doesn't sound like Alex is having an easy time of it either." Sophia turns to her younger colleague and says, "Trust me, Sam, I've been around long enough to know that you can't coddle these kids. They need boundaries. You'll learn quick or you won't last in this job."

Parental Meeting

Marylyn and Alicia, friends from college and now teachers at the same elementary school, are sitting having lunch together in a quiet spot in a park near their school. Alicia tells Marylyn that something upsetting happened recently at school with one of her pupils. "Can we keep this between us though?" Alicia asks, "I just need to get it off my chest." Marylyn nods. Alicia then tells Marylyn about a meeting she had the previous afternoon with the mother of one of pupils in her grade 4 class. The pupil in question is Jacob, new to the school this year. Even though they were now into the fourth week of the school year, he hadn't managed to make any friends and was still spending recess wandering around alone in the playground. He was also having trouble with the material she was teaching. The fact that he frequently arrived late at school and his homework wasn't done half the time, Alicia tells Marylyn, also didn't help. So, Alicia decided to arrange a meeting with Jacob's parents so she could find a way of working with them to get him on track at his new school. The pupil's mother arrived at the meeting alone and visibly flustered. Shortly after the two women sat down, the mother burst into tears, explaining that things at home were hard right now. Jacob's father had had drinking problem off and on for years and was going through a rough period. They were fighting all the time about him drinking too much

and as she was speaking pulled up her shirt sleeve and showed Alicia a bruise on her left wrist. The mother said very sincerely that she was sorry but with all this going on it was nearly impossible for her to help Jacob with his homework. When Alicia suggested the family get help, and offered to put them in contact with someone at the school who could find them the resources they need, the mother replied that they'd been to all kinds of therapists but it had been no help. Having lived with this situation for so long, she'd learned that the best thing to do was just to "ride it out". As the mother got up to leave she turned to Alicia and said, "Please, don't tell anyone about this. I can handle this on my own." Alicia had given her word. After hearing the story, Marylyn says to Alicia, "This is terrible! What are you going to do?" Alicia replies, "Well, I promised Jacob's mother that I wouldn't tell anyone, and that's what I plan to do."

Annex 2

Pilot Survey Summary Model Fit, Mean Measure and Separation by Scenario

PARENTAL MEETING

SUMMARY OF 18 MEASURED PERSONS								
			logit		INFIT		OUTFIT	
	total score	count	measure	model S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	5.2	9.0	0.22	0.92	0.96	-0.02	1.01	0.25
SEM	0.4	0.0	0.28	0.01	0.11	0.22	0.21	0.18
P.SD	1.5	0.0	1.16	0.05	0.45	0.92	0.85	0.75
S.SD	1.5	0.0	1.19	0.05	0.46	0.95	0.87	0.77
MAX.	7.0	9.0	1.62	1	2.17	1.87	3.41	2.17
MIN.	3.0	9.0	-1.58	0.86	0.47	-1.06	0.3	-0.52
Real RMSE 1.00 True S.D. 0.59				Separation 0.59			Person reliability 0.26	
Model RMSE 0.93 True S.D. 0.70				Separation 0.76			Person reliability 0.36	
S.E. of Person Mean = 0.28								

PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .35* SEM = 1.18

*Due to the narrow range of ethical sensitivity in the respondents.

SUMMARY OF 8 MEASURED items								
			logit		INFIT		OUTFIT	
	total score	count	measure	model S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	9.5	18.0	0.00	0.66	1.00	-0.05	1.01	0.4
SEM	1.7	0.0	0.63	0.06	0.10	0.42	0.13	0.38
P.SD	4.5	0.0	1.67	0.17	0.26	1.12	0.35	1.01
S.SD	4.8	0.0	1.79	0.18	0.27	1.20	0.37	1.08
MAX.	17.0	18.0	2.76	1.05	1.23	0.98	1.43	1.19
MIN.	2.0	18.0	-3.13	0.54	0.45	-2.57	0.39	-2.15
Real RMSE 0 .71 True S.D. 1.51				Separation 2.12			Item reliability 0.82	
Model RMSE 0.68 True S.D. 1.53				Separation 2.26			Item reliability 0.84	
S.E. of item Mean = 0.63								

MAXIMUM EXTREME SCORE: 1 ITEM 11.1%

SUMMARY OF 9 MEASURED (EXTREME AND NON-EXTREME) ITEMS								
			logit		INFIT		OUTFIT	
	total score	count	measure	model S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	10.4	18.0	-0.49	0.79				
SEM	1.8	0.0	0.74	0.14				
P.SD	5.0	0.0	2.10	0.40				
S.SD	45.2	0.0	2.22	0.43				
MAX.	18.0	18.0	2.76	1.83				
MIN.	2.0	18.0	-4.40	0.54				
Real RMSE 0 .91 True S.D. 1.89				Separation 2.08			Item reliability 0.81	
Model RMSE 0.88 True S.D. 1.90				Separation 2.16			Item reliability 0.82	
S.E. of item Mean = 0.74								

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.98

*Excellent separation of items.

FACULTY LOUNGE

SUMMARY OF 21 MEASURED (NON-EXTREME) PERSONS								
			logit		INFIT		OUTFIT	
	total score	count	measure	model S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	3.6	9.0	-0.66	0.86	.98	-.08	1.07	.15

SEM	0.3	0.0	0.21	0.01	.10	.25	.25	.21
P.SD	1.4	0.0	0.95	0.04	.45	1.11	1.10	.92
S.SD	1.4	0.0	0.97	0.04	,46	1.13	1.13	.94
MAX.	6.0	9.0	0.99	0.92	1.88	1.76	5.27	2.68
MIN.	2.0	9.0	-1.82	0.82	.41	-1.53	.30	-1.05
Real RMSE 0 .93 True S.D. 0.17				Separation .18			Item reliability 0.03	
Model RMSE 0.86 True S.D. 0.40				Separation .47			Item reliability 0.18	
S.E. of Person Mean = 0.21								

MINIMUM EXTREME SCORE: 1 PERSON 4.5%
*Due to the narrow range of ethical sensitivity in the respondents.

SUMMARY OF 22 MEASURED (EXTREME AND NON-EXTREME) PERSONS								
			logit		INFIT		OUTFIT	
	total score	count	measure	model S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	3.5	9.0	-0.82	0.90				
SEM	0.3	0.0	0.26	0.05				
P.SD	1.5	0.0	1.19	0.22				
S.SD	1.6	0.0	1.22	0.22				
MAX.	6.0	9.0	0.99	1.88				
MIN.	0.0	9.0	-4.24	0.82				
Real RMSE 0 .99 True S.D. 0.65				Separation .65			Item reliability 0.30	
Model RMSE 0.93 True S.D. 0.74				Separation .80			Item reliability 0.39	
S.E. of Person Mean = 0.26								

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .35
SEM = 1.24

SUMMARY OF 9 MEASURED (NON-EXTREME) ITEMS								
			logit		INFIT		OUTFIT	
	total score	count	measure	model S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	8.4	22.0	0.00	0.61	0.99	0.00	1.07	0.16
SEM	1.9	0.0	0.56	0.06	0.04	0.17	0.17	0.24
P.SD	5.3	0.0	1.57	0.18	0.12	0.49	0.47	0.69
S.SD	5.6	0.0	1.67	0.19	0.13	0.52	0.50	0.73
MAX.	17.0	22.0	2.73	1.04	1.18	0.82	2.16	1.26

MIN.	1.0	22.0	-2.36	0.48	0.86	-0.78	0.46	-0.69
Real RMSE 0 .65 True S.D. 1.43				Separation 2.19			Item reliability 0.83	
Model RMSE 0.63 True S.D. 1.44				Separation 2.28			Item reliability 0.84	
S.E. of item Mean = 0.56								

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.99

*Excellent separation of items.

Annex 3

Rasch Measure of Ethical Sensitivity by Country Codes as Dummy Variables

	RASCH measure of ethical sensitivity	
	Mean	S. D.
Australia	-.78	.66
USA	-.87	.84
Singapore	-.80	.44
Canada	-.57	.62