Moral reasoning assessment for Singapore secondary schools: A review

Lyndon Lim
Singapore University of Social Sciences, Singapore
Elaine Chapman
The University of Western Australia, Australia

Endeavours to assess moral reasoning in education have largely been via established but less contemporary measures, with recent measures developed more than a decade ago. Together with the call to go beyond assessing cognitive constructs, and the intended learning outcomes of the Singapore Ministry of Education Character and Citizenship Education curriculum that requires teachers to help students progress in their moral reasoning stages, there is a need for a measure that affords consistency when evaluating students' attainment of learning outcomes stipulated in the curriculum. Guided by Messick's unitary concept of validity, this paper reviewed existing measures of moral reasoning for suitability, and found that established measures presented varying degrees of tenability in assessing moral reasoning. Findings related to content appropriateness and group administrability yielded a paucity of measures applicable for large-scale assessment of moral reasoning in Singapore secondary schools. To address some of these issues, this review suggests the development of a fit-for-purpose measure.

Introduction

There has been an increasing focus to go beyond assessing cognitive constructs in Singapore (Ng, 2017). These include assessing attitudes and moral reasoning. Given the significance of moral reasoning and its assessment are key elements within the current Singapore *Character and Citizenship Education* (CCE) curriculum that aims to develop students into good and useful citizens (Singapore Ministry of Education [MOE], 2014, 2016), suitable assessment instruments could be deployed for use within classrooms of Singapore schools. This is particularly critical for secondary level (grades 7-12) classrooms as the CCE curriculum suggests that teachers assist students in reaching various levels of moral development by discussing moral dilemmas based on the clarify-sensitise-influence approach during curriculum time, and modelling how informed moral reasoning decisions should be made in the context of these dilemmas. While desirable given its personable nature, such an approach is resource and time intensive, and does not provide a practical means to track the moral development of students over time, which is an obligation, given that the CCE curriculum states that a student should progressively develop from being able to distinguish right from wrong at the primary level, having moral integrity at the secondary level, and having the moral courage to stand up for what is right at the pre-university level (MOE, 2014, 2016). Within the secondary level, the CCE curriculum also intends for students to progress through the conventional to post-conventional levels of moral reasoning based on Kohlberg's theory of moral development (Kohlberg, 1984); discussing dilemmas using the clarify-sensitise-influence approach may not present a sense of how students in a class have progressed unless multiple discussions take place over students' secondary school years and are meticulously logged.

Responding to the focus upon going beyond assessing cognitive constructs, this paper aims to review existing measures of moral reasoning within the literature that can be possibly deployed for use within Singapore classrooms. Suitably guided by the unitary concept of validity posited by Messick (1993, 1995) which is the standard for educational and psychological assessments, the subsequent sections discuss validity, particularly content appropriateness, reliability and manageability, of existing measures and their applicability for the assessment of moral reasoning on a broad-scale basis.

Types of measures

According to Palmer (2018), measures of moral reasoning can be categorised as: (i) production measures, which require a participant to construct a response to justify a decision, and (ii) recognition measures, which require a participant to recognise and select a response to support a decision.

Constructed response items, in general, provide more information on a respondent's mastery of the assessed construct. However, more time is required to complete a test with constructed response items, and scoring is more complex, with the introduction of rater subjectivities. Selected response items on the contrary may not provide as much information as constructed response items, but are generally more practical if a measure is to be used in large-scale settings. Despite the suggestion that a recognition measure might be inflated (Weber, 2018), these practical advantages favour a recognition measure though in areas as complex as moral reasoning, items should be developed in consultation with an expert panel, and trialled with a small sample of potential respondents, before they are deployed formally.

Existing measures of moral reasoning

This section focuses upon providing a review of existing measures to assess students' moral reasoning levels. The review, that draws from critiques, use cases and validity evidence, found that most of the established measures within the literature were developed some time ago with research, commentaries and use cases reported to date; there has been a dearth of novel and contemporary measures, given that the most recent measure was developed in the 2000s decade. As a presentation of similarities and differences, a brief description is provided for each of the measures reviewed. Advantages and disadvantages of each measure are highlighted along with recommendations of which measure, or the development of a new measure would be fit-for-purpose (for large-scale assessment, e.g. Singapore classrooms, and similar contexts).

Moral judgment interview

Developed by Lawrence Kohlberg who proposed the theory of moral development, the *Moral Judgment Interview* (MJI) can be considered the founding measure to assess moral reasoning. Kohlberg originally presented the MJI in his 1958 dissertation, which had been developed as an operationalisation of his theory of moral development (Kohlberg, 1958, 1984). As a production measure, the MJI administration involves interviewing students in a semi-structured format, with each interview lasting between 30 to 90 minutes depending on how the interviewee responds and how long the interviewer persists with specific questions in the interview. (Gibbs, Widaman & Colby, 1982). In the interview, the interviewer uses nine hypothetical moral dilemmas to determine which stage of moral reasoning the interviewee is applying when considering his/her response to each dilemma.

The moral dilemmas presented are fictional short stories that describe situations in which a person has to make a moral decision. The respondent is then asked a series of nine to twelve standardised and prescribed open-ended questions to elicit what they think the right course of action is, and, most importantly, why. Respondents are not asked what they would do or how they would act in that moral dilemma.

The MJI has three standard forms (Forms A, B and C), and stage scoring for each form is conducted on the basis of the *Standard Issue Scoring Manual.* In scoring the responses, what the participant thinks the agent in the dilemma should do is not important; it is the justification the participant offers for his/her choice that matters. In constructing the MJI, Kohlberg was not concerned with interviewees' judgments *per se* as his aim was to map their moral reasoning to the different stages within his model, with the assumption that the reasoning exhibited by respondents formed the basis for their judgments.

The scoring process for the MJI is complex and time-consuming because it relies on scorers' judgments of the stage that best fits a given interviewee's response to each of the dilemmas. In fact, Miller (2007) suggested that the MJI scoring might be the most complex scoring system in the field of psychology. Further, respondents who participate in 'poorly probed' interviews often need to be awarded 'guess' scores by scorers, because their reasoning can be unclear based on the responses received (Colby et al., 1983). Various scholars have questioned the reliability of the MJI based on these post-interview scoring practices. Cortese (1984), for example, highlighted some of the potential problems caused by 'guess' scores, which then factor into the overall MJI indices. 'Guess' scores can be related to how the interviewer conducts the MJI, and also, a lack of understanding or misunderstanding of the stage structures within the measure. Cortese (1984) further added even trained interviewers might disagree on the number of 'whys' an interviewee should be asked for each dilemma, and that an interviewer's choice of words would inevitably influence the interviewee's responses.

Despite the criticisms of the MJI scoring system, a small number of empirical evaluations have provided mixed support for the reliability and validity of the MJI. Colby et al. (1983), for example, reported high test-retest reliabilities of .98, .96 and .92 for the MJI Forms A, B and C, based on a one-month test-retest interval. In the same study, factor analyses and Cronbach's alphas also supported the dimensionality and internal consistency of the MJI (Form A alpha = .92, Form B alpha = .96 and Form C alpha = .90). Inter-rater reliability, however, was found to be less robust, ranging from .53 to 1.00 across the three forms, depending on the level of specificity of the judgments being made.

Despite the mixed positive evidence, the potential impact of interviewer subjectivity upon the scoring remains a significant stumbling block to the use of this measure. It is apparent that the validity and reliability of the MJI hinges not only on the interviewee's ability to articulate his or her reasoning, but also on the skill of the interviewer in eliciting interpretable responses. In a situation where an interviewee is less articulate than others, or in situations where interviewers have a tendency to under- or over-interpret conversations, the interpretability and hence, validity of the MJI scores would ultimately be compromised. Thus, while the MJI has been used extensively and has been reported to exhibit high levels of internal consistency and test-retest reliability (Colby et al., 1983; Gibbs et al., 2019), alternatives to mitigate practical issues such as the time required to do the test, reliance on highly trained interviewers, the complex scoring and coding system, and some of the inter-rater and validity issues associated with the test, are needed.

Defining issues test

According to Miller (2007), there are two alternatives to Kohlberg's MJI: (i) the *Defining Issues Test* (DIT), and (ii) the *Sociomoral Reflection Measure* (SRM). In contrast to the MJI, the DIT is a recognition measure of moral reasoning, originally developed by Rest (1979). The DIT is generally considered the primary alternative to Kohlberg's MJI for assessing stages in the original Kohlberg model. Though Rest (1979) stated he did not intend the DIT to be considered an optimal measure of moral reasoning, and encouraged further explorations of its properties in the original validation of the measure, the DIT is now a prominent alternative to the MJI (Gibbs et al., 1982, 2019). It is likely that the popularity of this measure stems in part from the fact that it is less time-consuming and less expensive to administer than the MJI.

There are six moral dilemmas in the original DIT and three in an abbreviated DIT (Elm & Weber, 1994; Weber & Elm, 2018). Respondents to the DIT read a series of moral dilemmas and, against each dilemma, rate 12 stage-related factors that could be considered in responding to the dilemma in terms of their importance, on a five-point scale (from of great to no importance). Respondents then select and rank four of these 12 items as most to least important.

According to Rest (1979), the DIT assumes that people at different developmental levels in their moral reasoning respond to each moral dilemma differently. In rating and ranking the stage-related considerations following the moral dilemmas, the DIT assumes that a respondent has used a specific stage or at most two adjacent stages of moral reasoning in making his or her response, noting that a subsequent stage is a reconstruction or transformation of the previous. Rest (1979) designed the DIT as a developmental measure of moral judgment by a two-stage process of preference and recognition. Thus, the DIT is not reliant on expressive verbal skills, unlike the MJI (although it does rely on the respondent's ability to read the dilemmas and ranking options).

A respondent's level of moral reasoning is represented by the P-score. While there have been concerns that score inflation or deflation might occur when respondents fake good (high) or bad (low), Rest (1979) cited various studies which suggested that faking would not impact DIT scores significantly. Further, the DIT cleverly includes meaningless statements that appear philosophical as 'foil' items. Scores of respondents who consistently select these statements would be disqualified as these suggest attempts to manipulate scores.

The validity and reliability of the DIT has been well-established (Christensen et al., 2016; Elm & Weber, 1994; Mudrack & Mason, 2021; Weber & Elm, 2018). For example, works by Rest et al. (1997b) suggest adequate to good internal reliability of the DIT (P-score), with Cronbach's alpha in the range of .76 for a 1979 composite sample (*N*=994) to .78 for a 1994 composite sample (*N*=932). In reviewing the empirical evidence relating to the validity of the DIT, Rest et al. (1997a) concluded that support for the measure was strong, based on evidence garnered across seven construct validity criteria: (i) differentiation of P-scores between age-education groups; (ii) longitudinal gains (in P-scores); (iii) correlation with cognitive capacity measures; (iv) sensitivity to moral education interventions; (v) correlation with behaviour and professional decision making; (vi) relations with political choice and attitude; and (vii) "fakeability" studies.

Despite the favourable empirical evidence cited, two significant criticisms of the DIT have appeared within the literature on moral reasoning measurement: (i) the issue of using a quantitative measure to describe a qualitative theoretical framework, and (ii) using the P-score as a reflection of an individual's moral development.

In responding to the qualitative-quantitative criticism, Rest and colleagues asserted that the DIT is developmental and not evaluative, and DIT scores were not intended to place an individual definitively in a particular stage. Rather, the DIT intended to measure the "extent of and under what conditions does a person manifest particular stages of thinking" (Rest et al., 1997b, p.499). In his study of another one hundred studies and in constructing the DIT, Rest (1979) concluded that moral judgment is developmental, and a major source of variation other than age is social experience. Hence, while the DIT could be considered for use in Singapore schools given that its intent was to measure the extent of how much an individual was at a particular stage, the interaction of social experience and familiarity, and how subjects responded to items presented the DIT as less appropriate; the dilemmas in the DIT seem non-familiar and dated as daily experiences to students at the secondary level or equivalent in Singapore.

Defining issues test 2

In response to problems related to the DIT (e.g., the P-score that does not consider non post-conventional ranking, dated dilemmas, potential respondent "fakeability", and group administrability), a revised version, the Defining Issues Test 2 (DIT2) was published (Rest et al., 1997a). Purported to measure moral reasoning based on Kohlberg's original model and parallel to the DIT in construction, the DIT2 is a recognition measure which provides quantitative scores based on the test-taker's responses to five hypothetical moral dilemmas (Rest et al., 1999b).

Similar in nature to the DIT but shorter and with clearer instructions, each of the five hypothetical moral dilemmas in the DIT2 is followed by 12 issues that could be considered in resolving the dilemma. Participants are asked to indicate a decision in each dilemma based on a seven-point scale ranging from strongly favour to strongly disfavour; they then indicate which of the listed issues are most important to their decision using a five-point scale ranging from great to no importance. Responses are scored to quantify, based on the schema theory by Rest et al. (1999a) resulting in an index known as the N2-score.

Owing to its comparatively limited history, the DIT2 has a less extensive empirical evidence base to support it than the original DIT. Rest et al. (1999b), and later Thoma and Dong (2014) have, however, reviewed studies and validated the DIT2 based on similar criteria to the seven used to validate the original DIT. They concluded that the empirical evidence largely supported the DIT2 as a valid measure of moral reasoning. Thus, in comparison to the MJI, the DIT/DIT2 is considered by many to have a stronger evidence base.

Despite evidence to support the validity of the DIT and later the DIT2 (Choi et al., 2020), criticisms remain. An early critique was proffered by Kay (1982), who asserted that Rest and colleagues used correlational designs confounded by extraneous variables in their cross-sectional and longitudinal studies, and the failure to isolate study variables was a severe limitation of the studies; this could result in P-scores being confounded by other variables. While this assertion undermines the notion of the DIT as a measure of moral development, it is important to note that DIT developers had never claimed that it was intended to measure moral development in its entirety.

A more recent critique of the DIT and DIT2 was by Curzer et al. (2014), in their attempt to develop an alternative to the DIT called the *Sphere-Specific Moral Reasoning and Theory Survey*. This critique drew a strong response from DIT researchers. Thoma et al. (2016) stated categorically that the comments by Curzer and colleagues were untenable as they did not rely on empirical evidence and had misunderstood the DIT and its corresponding models.

Ethical reasoning inventory

The *Ethical Reasoning Inventory* (ERI), developed and validated by Page and Bode (1980) is similar in content and foci to the DIT and DIT2 but simpler to administer and score. The ERI was developed partly in response to the internal consistency issue reported for the initial DIT, in comparison to those reported previously for the MJI (e.g., Cronbach's alpha at the dilemma level for the DIT was .65, compared to .89 for the MJI). The ERI requires participants to respond to six dilemmas similar to those used by Kohlberg in the MJI. For each dilemma, participants first select one of two 'action' options. Based on their selected response, participants turn to the relevant page and select, out of six options corresponding to Kohlberg's six stages, an option that best represents their reasoning for choosing the action they did. Thus, participants need to select two options for each dilemma presented.

The ERI has not been validated as extensively as the DIT and DIT2 though comparisons between different measures of moral reasoning have suggested that the ERI could be more reliable than the MJI and DIT. For example, in a study by Page and Bode (1980) in which the MJI, DIT and ERI were administered to a sample of college freshmen and sophomores (*N*=92), the correlation between scores of the MJI and ERI was slightly higher (*r*=.54) than that between the MJI and DIT (*r*=.50). Further, the coefficient of stability of the ERI under test-retest conditions of college students (*N*=51) was .69 with an interval of 10 days; this was slightly higher than that of the DIT at .65 (*N*=47) with an interval of 18 days. Thomas (1986) stated that a higher coefficient of stability points to a more consistent assessment technique over time. Hence, the ERI in this instance can be considered more consistent. It is noteworthy, however, that the sample of college students used for comparison was not the same.

While Page and Bode (1980) used Pearson correlation coefficients to demonstrate superior consistency of the ERI, there have been limited empirical validation studies. The dilemmas used were similar to those used in Kohlberg's MJI, but there was no mention that the 'action' options and the latter six options were deemed content-appropriate by experts. There was also limited discussion establishing the internal factorial structure of the ERI.

Nevertheless, Bode and Page (1979) stated that the ERI possesses reliability, construct validity and can be used with subjects aged as young as 14 based on their definition of validity. In addition, their "fakeability" studies on 174 college students suggested that respondents were unable to fake upwards though they could fake downwards significantly (Page & Bode, 1979). While the ERI is more easily group-administrable compared with the MJI, Rest et al. (1997a) stated that higher alpha values may not mean a measure is better than another, as opposed to Page and Bodes' (1980) earlier claim that the ERI had a better reliability coefficient of internal consistency. Given the lack of literature, including contemporary ones, on the ERI and the existing literature on the DIT and DIT2, it cannot be concluded that the ERI is a superior measure to the DIT for assessing moral reasoning development.

Sociomoral measures

More recently developed measures to assess moral reasoning based on Kohlberg's stages of moral development include the *Sociomoral Reflection Measure* (SRM) developed by Gibbs et al. (1982) and its derivatives (i.e., the *Sociomoral Reflection Objective Measure*, or SROM, developed by Gibbs et al. (1984); the *SROM - Short Form*, or SROM-SF, developed by Basinger and Gibbs (1987); the *Sociomoral Reflection Measure - Short Form*, or SRM-SF, developed by Gibbs et al. (1992); and the most recent Sociomoral Reflection Measure – Short Form Objective, or SRM-SFO, developed by Brugman et al. (2007); Gibbs et al. (2019). As with the DIT, the SRM has been demonstrated to have favourable psychometric properties, though the DIT has been reported to exhibit a lower correlation with MJI scores than the SRM (Palmer, 2018). The SRM also requires respondents to construct their responses as opposed to that of the DIT.

The initial SRM was an attempt to make the MJI more group-administrable in light of criticisms of the MJI which requires a substantial investment of time, effort, and cost for its effective use (e.g., interviewers had to attend a five to ten-day workshop just to learn about the MJI scoring structure) (Gibbs et al., 1982). In addition, Gibbs et al. (1982) cited evidence that the primary alternative to the MJI, the DIT, did not correlate adequately with the MJI when chronological age was partialed out. Hence, Gibbs and colleagues sought to develop the SRM, an MJI equivalent that is more feasible in terms of administration.

Similar to the MJI, the SRM is a production measure and assesses justificatory moral judgment that lasts about 30 minutes less than what the MJI requires. In the SRM, respondents reflect and express their thoughts on moral dilemmas similar to those used in the MJI. A key difference is that probe questions in the SRM were modified from the MJI so that these questions may be more consistently efficacious, accommodate sufficient scorable responses and hence minimise "guess" scores, an issue the MJI faces.

Scorers/raters have to undergo training before scoring the SRM. Nonetheless, the training can be conducted in a minimum of six hours which is relatively more manageable than that associated with the MJI training, which lasts a minimum of five days. It is noteworthy that, however, requiring classroom teachers to undergo six hours of training to be scorers/raters would be less preferred compared with no training or if additional resources were deployed to undergo the training.

In validating the SRM, Gibbs et al. (1982) involved a sample of 107 subjects (59 female), aged from 12 to 22 (mean age = 15.5 years old) and studied four kinds of reliability: (1) inter-rater; (2) test-retest; (3) parallel form; and (4) internal consistency. Though the sample size was small, it was concluded that the SRM had acceptable inter-rater reliability, reliability and internal consistency. Gibbs and colleagues also found that the SRM covaried significantly with expected variables (age, grade and social economic status, or SES) and did not covary with sex, suggesting that the measure was not gender biased. With these sources of evidence supporting the validity of the SRM (Colby et al., 1983), various derivatives were developed to further mitigate against practical limitations associated with administration (Gibbs et al., 1984).

A derivative of the SRM, the SROM was developed as a recognition measure and required about 20 minutes less to administer and less time to score (Basinger & Gibbs, 1987). In the SROM, participants have to rate and rank statements similar to the DIT and DIT2 after reading moral dilemmas also used in the MJI and SRM. Respondents have to then complete 16 multiple-choice arrays. As an alternative to the MJI, most of the components of this measure were designed to correspond to Kohlberg's stages of moral development.

Test-retest studies by Basinger and Gibbs (1987) on the SROM conducted with a two-week interval yielded a correlation of .82 (.76 with age partialed out); the lowest correlation of .70 was from the seventh graders but this was considered acceptable. Cronbach's alpha (.84) suggested adequate internal consistency though the sample was small. The SROM also correlated adequately with the SRM ($r(81) = .73$, $p < .001$) in a study with 82 subjects aged 11 to 22 (mean age = 14.5 years old), and the MJI ($r(21) = .66$, *p* < .001), though this result should be interpreted with caution as only 23 subjects aged 13 to 41 (mean age = 20.1 years old) were involved.

To gather further evidence on the SROM's construct validity, Gibbs et al. (1984) studied correlations between the SROM and variables including age, grade, IQ, SES and social desirability. They found that correlations with age and grade were significant and that with IQ was significant with a larger sample. Though the correlation with grade was in the .60s, an upward progression was observed in the mean SROM and the SRM scores; this suggested that Kohlberg's theory could be demonstrated by grade levels. Nonetheless, Gibbs et al. (1984) expressed that the SROM could not distinguish delinquents from non-delinquents when IQ was not partialed out. Further, they conceded based on the evidence they had collected that the SROM might not be applicable to all adolescent and age levels (e.g., sixth graders) as reading literacy was a prerequisite.

In an attempt to further shorten the SROM, Basinger and Gibbs (1987) developed the SROM-SF, a group-administrable and purportedly more objective measure that involves inferences by interviewers and is hence easier to score. Respondents have to complete a questionnaire comprising two moral dilemmas and 48 moral reasoning justifications in the SROM-SF. The SROM-SF excludes items that were more verbally complex in the SROM. Basinger and Gibbs (1987) found the SROM-SF reliable and valid specifically with eleventh graders and it required about 20 minutes less than the SROM for administration. Nonetheless, as with the SROM, there was a lack of evidence to conclude that the SROM-SF would be applicable to sixth graders and juvenile delinquent adolescents.

Following the development of the SROM-SF, Gibbs et al. (1992) developed the SRM-SF in an effort to shorten and simplify the initial more complex SRM for efficiency. Similar to the SRM, the SRM-SF is a production measure anchored on Kohlberg's theory. Though shorter and initially touted by Gibbs and colleagues as more group-administrable than the SRM, the extent of group administration of the SRM-SF is questionable beyond the classroom context given that interviews still have to be conducted, transcribed, analysed and scored (Brugman et al., 2007).

Respondents for the SRM-SF are required to complete an 11-item questionnaire by circling, for each item, one of the options "very important/ important/ not important" and explaining in writing why they chose that option. Instead of longer moral dilemmas, short scenarios were used.

Correlation analyses have been performed to evaluate the SRM-SF though with a limited sample size. It was found that there was high inter-rater reliability, Cronbach's alpha and hence, acceptable reliability (Brugman et al., 2007).

Of the four measures on sociomoral assessment, the SRM-SF has been more widely used and can also be applied to a wider age group; it has also stronger evidence for construct validity and reliability compared with the other three (Bock, 2008). Besides, Palmer (2018) stated that the SROM and SROM-SF have proven to have relatively more limited reliability and validity. Bock (2008) highlighted that the SRM-SF is unique in that it uses moral behaviours (vignettes) derived from Kohlbergian moral dilemmas instead of the usual lengthy moral dilemmas. Nonetheless, participants are still required to reason in writing their choice of a particular moral behaviour; this could explain why IQ correlated positively with the SRMS-SF as a respondent with a higher IQ score would generally be more articulate in reading and writing. While Bock (2008) stated that it has very good to excellent psychometric properties, it does not include Kohlberg's stages five and six as these would require higher verbal abilities. Both the SRM and SRM-SF have been tested beyond the American context and proven to be applicable though the SRM-SF is easier to score (Ferguson et al., 1994; Gibbs et al., 2019; Nilsson et al., 1991). Despite these results, it should be noted that the scenarios presented in the SRM-SF and the moral dilemmas of the SRM may not be familiar in the Singapore context, given the characteristics of the sample used to validate the SRM and SRM-SF.

In a further bid to reduce administration and coding time, Brugman et al. (2007) developed the SRM-SFO. The SRM-SFO combines both the SRM-SF and SROM-SF, noting the shortcomings of both (e.g., the SROM-SF poses problems for younger respondents especially those who have reading difficulties given the moral dilemmas are used in the measure, and for the SRM-SF, being a production measure, it requires fundamental writing skills and likely pose challenges for administration beyond the size of a classroom). The SRM-SFO comprises ten sets of questions that participants rate and rank in about 15 minutes, much lesser time required than its predecessors.

In validating the SRM-SFO, Brugman et al. (2007) focused on comparing the scores of delinquents and non-delinquents and found that the moral maturity of non-delinquents (SRM_score_mean = 281, *SD* = 40.4, range: 170-367) was unexpectedly slightly lower than that of delinquents; this could be explained by the older age of the delinquent group (SRM_score_mean = 294, *SD* = 26.8, range: 225-369). This unexpected result was confirmed with an updated sample comprising 107 non-delinquents (*M* = 14.0, *SD* = 1.0) and 45 delinquents (*M* = 15.1, *SD* = .71). Surprisingly, these results are contrary to those of the SRM-SF (that delinquents scored lower than non-delinquents).

Further to the comparison between delinquents and non-delinquents, Brugman and colleagues performed a confirmatory factory analysis on the SRM-SFO and a multi-group confirmatory factor analysis to compare the delinquent and non-delinquent groups. The multi-group analysis had acceptable fit $\chi 2$ (70, N = 152) = 75.96, $p < .29$; CFI = .97; RMSEA = .014) though some of the item loadings across both models were small.

Given its more recent development and the fewer validity and reliability studies, the SRM-SFO has not been used as widely as the SRM to date, and hence, may not be as applicable for the Singapore context. Further, a few items within the SRM-SFO may not be relatable to younger secondary school students, particularly those with language issues (e.g., responding to the item "people are not allowed to take away things that belong to others because living in society means accepting obligations and not only benefits"). Nonetheless, Brugman et al. (2007) concluded that the SRM-SFO holds promise as a measure of moral reasoning for adolescents given its format, ease of administration and initial acceptable level of reliability.

Test of moral values

An extensive search yielded only one measure that has been used to assess moral reasoning in the Singapore context. Soh (1987) developed the *Test of Moral Values* (TMV) for Singapore students in response to a call for a more organised moral education in Singapore. Unlike the previously discussed moral reasoning measures that are anchored on established theoretical models, the TMV comprises 24 items each anchored on a different moral value recommended in a 1979 Singapore Government report on moral education by the then Communications Minister and Acting Minister for Culture, Ong Teng Cheong.

Each option in each item within the TMV corresponds to a category (i.e., self, social/peer influence or moral value). Validation of TMV has been very minimal and lacked rigour. Hence, the TMV, remains non-validated for use though the number of categories appear favourable in that respondents might be more amenable to responding to three categories, as opposed to more categories.

Discussion

Despite the substantial body of research evidence which supports the validity of the measures discussed thus far (Table 1), and the paucity of more recent measures, for several reasons, none would be most suitable for assessing moral reasoning development on a broad-scale basis within the Singapore CCE secondary curriculum.

First, while the MJI does provide an extensive evaluation of students' moral reasoning, this test is individually administered, and extremely time-consuming and complex to score (Miller, 2007). Teachers in schools who are charged with assessing, in some cases, hundreds of students concurrently, would find using the MJI prohibitive. Teachers would also have to go through extensive training on the MJI scoring protocols and systems to minimise disparities. The same issues would also apply to other production measures such as the SRM and SRM-SF, and for these reasons, the latter measures are also not suitable for use within this context.

Second, while various established measures have been used with children across ages and cultures, students in the Singapore secondary context may not relate to the scenarios presented within these measures. Most focus on issues that will be unfamiliar to students in their day-to-day lives. As an example, while time efficient, the SRM-SFO comprises items that younger secondary school students may have difficulty relating with, particularly for students who have language difficulties (e.g., responding to the item "people are not allowed to take away things that belong to others because living in society means accepting obligations and not only benefits").

Third, some hypothetical moral dilemmas used in the measures discussed may be inappropriate for students in the Singapore context given their complexity and that they are sometimes lengthy and difficult to comprehend. Further, terms used such as "*habeas corpus*" in the DIT2 would be considered unfamiliar to secondary school students based on the Common European Framework of Reference for English language teachers. For example, Gibbs et al. (1992) conceded that the SROM-SF includes fairly sophisticated dilemmas and has a format that sometimes makes it confusing for younger students. Shorter alternatives such as the SRM-SF, that come without lengthy moral dilemmas, may be more accessible to students at the secondary level but this form requires students to justify their selected responses in writing. This aspect could then introduce construct-irrelevant sources of variance, via differences in writing ability.

Other measures reviewed were either not supported by strong validation evidence (e.g., the TMV) or suffered from similar issues to those identified with the MJI and DIT/DIT2.

Table 1: Chronological summary of moral reasoning measures within this review

Instrument	Type of measure	No. parallel test forms	What participants have to do	Psychometric properties
Kohlberg (1958): Moral judgment interview (MJI)	Production	Two (Form A and B)	Construct response verbally or in writing to an interview based on a minimum of 21 probing questions per dilemma. There are 3 dilemmas per test form.	Acceptable levels of reliability and validity
Rest (1979): Defining issues test (DIT)	Recognition	One. An abbreviated form of 3 dilemmas can be used.	Select response (i.e., rate and rank) to 12 issue statements related to the initial decision to a moral dilemma. There are six dilemmas (three are Kohlbergian dilemmas).	Acceptable levels of reliability and validity
Page & Bode (1980): Ethical reasoning inventory (ERI)	Recognition	One	Select response (i.e., multiple-choice) to six moral dilemmas. There are 6 options each corresponding to a 'yes' or 'no' option for each dilemma.	Acceptable levels of reliability and validity
Gibbs et al. (1982): Sociomoral reflection measure (SRM)	Production	Two (Form A and B)	Construct responses to two moral dilemmas based on eight probing questions.	Acceptable levels of reliability and validity
Gibbs et al. (1984): Sociomoral reflection objective meas-ure (SROM)	Recognition	One (modelled after Form A of the SRM)	Select response (i.e., rate and rank) 16 multiple-choice arrays.	No acceptable validity and reliability for 6th graders and juvenile delinquents (Basinger & Gibbs, 1987)
Basinger & Gibbs (1987): Sociomoral reflection objective measure - Short form (SROM-SF)	Recognition	One	Select response (i.e., rate and rank) to two moral dilemmas.	No acceptable validity and reliability for 6th graders and juvenile delinquents (Basinger & Gibbs, 1987)

Gibbs et al. (1992): Sociomoral reflection reasure - Short form (SRM-SF)	Production	One	Construct responses to five short moral vignettes based on 11 questions.	Acceptable levels of reliability and validity
Rest et al. (1999b) DIT2	Recognition	One	Select response (i.e., rate and rank) to 12 issue statements related to the initial decision to a moral dilemma; 5 moral dilemmas.	Acceptable levels of reliability and validity
Brugman et al. (2007): Sociomoral reflection measure - Short form objective	Recognition	One	Select response to 10 dilemma free items.	Inconclusive

Conclusion and recommendations

The Organisation for Economic Co-operation and Development has consistently considered the Singapore education system as one of the most successful in the world. This has also been echoed by various sources, some suggesting the possibility of other countries learning from Singapore schools (Simonds, 2018). Nonetheless, some have called for the inclusion of standardised moral measures into the indices considered when ranking countries (Tan, 2015). This, the call to go beyond assessing cognitive constructs in Singapore, and the intended learning outcomes of the Singapore CCE curriculum, clearly present a need for a measure suitable for practical use in Singapore schools. While the Singapore CCE curriculum recommends a variety of assessment practices for teachers involved in CCE, to date, however, there are no standardised measures available to schools for the assessment of moral reasoning in students in a practical setting and hence, there has been no practical way to track students' moral reasoning stages and development. This introduces a potential problem in the lack of consistency with which schools may apply and evaluate students' attainment of the learning outcomes stipulated in the CCE.

Despite suitability issues of deploying established measures to assess moral reasoning in Singapore schools, this review reinforces the notion that it was definitively possible to assess such a construct (i.e., moral reasoning) via a measure anchored on Kohlberg's theory, a theory that is also mentioned within the CCE curriculum. It also highlighted qualities desirable and practical for the Singapore context of a such a measure: (1) recognition instead of production measures would be preferable; (2) short instead of long moral dilemmas (vignettes) familiar to the Singapore student should be used where possible; (3) simpler and less confusing response options (e.g., the ERI) should be used; (4) the measure should be short and group-administrable (ideally about 30 minutes for a sitting – the equivalent of one class period in a day); (5) there should be little or no need for classroom or CCE teachers to undergo scorer/rater training; and (6) moral reasoning progression demonstrated by such a measure should be anchored on Kohlberg's theory of moral development that is stipulated in the CCE curriculum. In light of this, it would be worthwhile to develop a measure suitable for assessing Singapore secondary students' moral reasoning within the CCE curriculum. This measure could also be extended to contexts within which Kohlberg's established theory of moral development is anchored upon.

References

Basinger, K. S. & Gibbs, J. C. (1987). Validation of the Sociomoral Reflection Objective Measure - Short Form. *Psychological Reports,* 61, 139-146. https://doi.org/10.2466/pr0.1987.61.1.139

Bock, T. (2008). Sociomoral Reflection Measure. In F. C. Power, R. J. Nuzzi, D. Narvaez, D. K. Lapsley & T. C. Hunt, (Eds.), *Moral education: A handbook, volume 2: M-Z* (pp. 424-426). Praeger. https://psycnet.apa.org/record/2008-01808-000

Bode, J. & Page, R. (1979). Further validation of the Ethical Reasoning Inventory. *Psychological Reports,* 45(3), 985-986. https://doi.org/10.2466/pr0.1979.45.3.985

Brugman, D., Basinger, K. S. & Gibbs, J. C. (2007). Measuring adolescents' moral judgment: An evaluation of the Sociomoral Reflection Measure – Short Form Objective (SRM-SFO). Presented at *Symposium: Cross-cultural research on moral reasoning,* International Council of Psychologists, San Diego, USA, 11-14 August. https://www.academia.edu/4594794/Measuring_Adolescents_Moral_Judgment_An_Evaluation_of_the_Sociomoral_Reflection_ Measure_-_Short_Form_Objective_SRM-SFO_

Choi, Y., Han, H., Bankhead, M. & Thoma, S. J. (2020). Validity study using factor analyses on the Defining Issues Test-2 in undergraduate populations. *PLoS ONE,* 15(8), 1-18. https://doi.org/10.1371/journal.pone.0238110

Christensen, A. L., Cote, J. & Latham, C.K. (2016). Insights regarding the applicability of the Defining Issues Test to advance ethics research with accounting students: A meta-analytic review. *Journal of Business Ethics,* 133, 141-163. https://doi.org/10.1007/s10551-014-2349-7

Colby, A., Kohlberg, L., Gibbs, J. C., Lieberman, M., Fischer, K. & Saltzstein, H. D. (1983). A longitudinal study of moral judgment. *Monographs of the Society for Research in Child Development,* 48(1/2), 1-124. https://doi.org/10.2307/1165935

Cortese, A. J. (1984). Standard issue scoring of moral reasoning: A critique. *Merrill-Palmer Quarterly,* 30(3), 227-246. https://www.jstor.org/stable/23086098

Curzer, H. J., Sattler, S., DuPree, D. G. & Smith-Genthôs, K. R. (2014). Do ethics classes teach ethics? *Theory and Research in Education,* 12(3), 366-382. https://doi.org/10.1177/1477878514545209

Elm, D. R. & Weber, J. (1994). Measuring moral judgment: The Moral Judgment Interview or the Defining Issues Test? *Journal of Business Ethics,* 13(5), 341-355. https://www.jstor.org/stable/25072538

Ferguson, N., McLernon, F. & Cairns, E. (1994). The Sociomoral Reflection Measure – Short Form: An examination of its reliability and validity in a Northern Irish setting. *British Journal of Educational Psychology,* 64(3), 483-489. https://doi.org/10.1111/j.2044-8279.1994.tb01119.x

Gibbs, J. C., Widaman, K. F. & Colby, A. (1982). Construction and validation of a simplified group-administrable equivalent to the Moral Judgment Interview. *Child Development,* 53(4), 895-910. https://doi.org/10.2307/1129126

Gibbs, J. C., Arnold, K. D., Morgan, R. L., Schwartz, E. S., Gavaghan, M. P. & Tappan, M. B. (1984). Construction and validation of a multiple-choice measure of moral reasoning. *Child Development,* 55(2), 527-536. https://doi.org/10.2307/1129963

Gibbs, J. C., Basinger, K. S. & Fuller, D. (1992). *Moral maturity: Measuring the development of sociomoral reflection*. Erlbaum. https://www.routledge.com/Moral-Maturity-Measuring-the-Development-of-Sociomoral-Reflection/Gibbs-Basinger-Fuller-Fuller/p/book/9781138976436

Gibbs, J. C., Basinger, K. S., McDonald, R. & Lee, D. (2019). Moral judgment maturity: From clinical to standard measures. In M. W. Gallagher & S. J. Lopez (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 333-346). American Psychological Association. https://doi.org/10.1037/0000138-021

Kay, S. R. (1982). Kohlberg's theory of moral development: Critical analysis of validation studies with the Defining Issues Test. *International Journal of Psychology,* 17(1-4), 27-42. https://doi.org/10.1080/00207598208247430

Kohlberg, L. (1958). *The development of modes of moral thinking and choice in the years 10 to 16.* PhD thesis, The University of Chicago, USA. http://search.proquest.com/docview/301935075

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

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, Volume 1993, issue 2, p. i-18. https://doi.org/10.1002/j.2333-8504.1993.tb01562.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50(9), 741-749. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Miller, S. A. (2007). Social development. In *Developmental research methods* (pp. 269-312). SAGE. https://doi.org/10.4135/9781412983891.n13

Mudrack, P. E. & Mason, E. S. (2021). Vignette themes and moral reasoning in business contexts: The case for the Defining Issues Test. *Journal of Business Ethics*. Online first. https://doi.org/10.1007/s10551-021-04944-8

Nilsson, I., Crafoord, J., Hedengren, M. & Ekehammar, B. (1991). The Sociomoral Reflection Measure: Applicability to Swedish children and adolescents. *Scandinavian Journal of Psychology,* 32, 48-56. https://doi.org/10.1111/j.1467-9450.1991.tb00852.x

Ng, C. M. (2017,). *Speech by Minister Ng Chee Meng for MOE (Schools) at the Committee of Supply Debate*. 7 March [not found 3 Nov 2021] https://www.gov.sg/microsites/budget2017/press-room/news/content/speech-by-minister-ng-chee-meng-for-moe-schools-at-the-committee-of-supply-debate

Page, R. & Bode, J. (1979). Degree of susceptibility to faking of the Ethical Reasoning Inventory. *The Journal of Educational Research,* 79(6), 355-356. https://doi.org/10.1080/00220671.1979.10885190

Page, R. & Bode, J. (1980). Comparison of measures of moral reasoning and development of a new objective measure. *Educational and Psychological Measurement,* 40(2), 317-329. https://doi.org/10.1177/001316448004000206

Palmer, E. J. (2018). Moral reasoning assessment. In R. J. R. Levesque (Ed.), *Encyclopaedia of adolescence.* Springer. https://doi.org/10.1007/978-3-319-33228-4_10

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

Rest, J., Thoma, S. J. & Edwards, L. (1997a). Designing and validating a measure of moral judgment: Stage preference and stage consistency approaches. *Journal of Educational Psychology,* 89(1), 5-28. https://psycnet.apa.org/doi/10.1037/0022-0663.89.1.5

Rest, J., Thoma, S. J., Narvaez, D. & Bebeau, M. J. (1997b). Alchemy and beyond: Indexing the defining issues test. *Journal of Educational Psychology,* 89(3), 498-507. https://psycnet.apa.org/doi/10.1037/0022-0663.89.3.498

Rest, J., Narvaez, D., Bebeau, M. J. & Thoma, S. J. (1999a). A neo-Kohlbergian approach: The DIT and schema theory. *Educational Psychology Review,* 11(4), 291-324. https://doi.org/10.1023/A:1022053215271

Rest, J., Narvaez, D., Thoma, S. J. & Bebeau, M. J. (1999b). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology,* 91(4), 644-659. https://doi.org/10.1037/0022-0663.91.4.644

Simonds, D. (2018). What other countries can learn from Singapore's schools. *The Economist*, 1 September. https://www.economist.com/leaders/2018/08/30/what-other-countries-can-learn-from-singapores-schools

Singapore Ministry of Education (2014). *2014 Syllabus, Character and Citizenship Education (Secondary).* https://www.moe.gov.sg/-/media/files/programmes/2014-character-citizenship-education-secondary.pdf?la=en&hash=45889182909A7AF1889080B6C2A50E03261764E0

Singapore Ministry of Education (2016). *Character and citizenship syllabus (Pre-university).* https://www.moe.gov.sg/-/media/files/programmes/character_and_citizenship_education_preu_syllabus.pdf?la=en&hash=F4B55203C3F0B328B45A6E703E7529AC2CC0FA0B

Soh, K. C. (1987). Test of moral values: Its development and try-out. *Singapore Journal of Education,* 8(2), 75-79. https://doi.org/10.1080/02188798708547626

Tan, T. H. (2015). Is it time for a new approach to education in Singapore? Towards education for a flourishing life (THF Lecture Series 2015). The Head Foundation. https://headfoundation.org/wp-content/uploads/2020/11/thf-papers_Is-it-time-for-a-new-approach-to-education-in-Singapore_-Towards-education-for-a-flourishing-life.pdf

Thoma, S. J., Bebeau, M. J. & Narvaez, D. (2016). How not to evaluate a psychological measure: Rebuttal to criticism of the Defining Issues Test of moral judgment development by Curzer and colleagues. *Theory and Research in Education,* 14(2), 241-249. https://doi.org/10.1177/1477878516635365

Thomas, R. M. (1986). Assessing moral development. *International Journal of Educational Research,* 10(4), 347-476. https://doi.org/10.1016/0883-0355(86)90001-7

Weber, J. & Elm, D. R. (2018). Exploring and comparing cognitive moral reasoning of millennials and across multiple generations. *Business and Society Review,* 123(3), 415-458. https://doi.org/10.1111/basr.12151

Weber, J. (2018). Does it matter how one assesses moral reasoning? Differences (biases) in the recognition versus formulation tasks. *Business & Society,* 57(7), 1440-1464. https://doi.org/10.1177/0007650316675611

Lyndon Lim *EdD* (corresponding author) is a Senior Lecturer in the Teaching & Learning Centre at the Singapore University of Social Sciences. Along with his prior research and work experience at the Singapore Ministry of Education and Singapore Examinations & Assessment Board, his research and publications focus on assessment and evaluation, psychometrics, and the social psychology of education.
ORCID: https://orcid.org/0000-0002-8199-5761
Email: lyndonlimjk@suss.edu.sg

Elaine Chapman *PhD* is an Associate Professor and Deputy Head of School – Research in the Graduate School of Education at The University of Western Australia. Her background is in psychology, but she has always had an interest in applying knowledge from psychology to education. Her general research interests lie in the areas of applied social and educational psychology, educational assessment, and research methods.
ORCID: https:// orcid.org/0000-0001-5861-1179
Email: elaine.chapman@uwa.edu.au

Please cite as: Lim, L. & Chapman, E. (2021). Moral reasoning assessment for Singapore secondary schools: A review. *Issues in Educational Research*, 31(4), 1121-1137. http://www.iier.org.au/iier31/lim.pdf