Development and Preliminary Validation of the Moral Reasoning Questionnaire for Secondary School Students

Lyndon Lim1,2 and Elaine Chapman3

Abstract

Moral reasoning is a key element of the Singapore Ministry of Education’s Character and Citizenship Education curriculum, and educators in Singapore are currently encouraged to facilitate growth in students’ moral reasoning based on the Stages of Moral Development established by Lawrence Kohlberg. To monitor students’ progress in this area, however, a validated test of moral reasoning is needed. A review of existing moral reasoning assessments indicated that none provide a practical means by which large numbers of students in Singapore classrooms can be assessed in this area. In this study, the Moral Reasoning Questionnaire (MRQ) was developed and validated to assess moral reasoning development in Singapore secondary classrooms. The MRQ presents moral dilemmas that integrate prosocial, anti-social, and social pressure elements within situations that would be familiar to Singapore secondary (grades 7–12, aged between 12 and 18) students. Evidence supported the reliability and validity of the MRQ. The MRQ could be applicable as a practical means to assess moral reasoning on a broad-scale basis.

Keywords

moral reasoning, assessment, secondary education, validation

Introduction

Recent years have seen a revival of interest in the subject of moral education across the globe. Debates on whether and how moral development and values are instructed in schools in particular have appeared in scholarly discourse emanating from both Western and Asian countries (Anzai & Matsuzawa, 2014; Bleazby, 2020; Lee & Park, 2019; Meindl et al., 2018; O’Flaherty et al., 2018). Whilst the questions of whether and how such education should be done remain highly contentious, it is clear from these sources that moral education is very much a current issue in today’s education systems internationally.

Singapore has been no exception in this global trend. The Singapore Ministry of Education (MOE) has specified in part that every Singaporean who completes formal education in Singapore should possess a moral compass, and that they be confident persons with a “strong sense of right and wrong” and are “discerning in judgment” (Singapore Ministry of Education, 2021). This is reflected across all of the MOE’s Character and Citizenship Education (CCE) syllabuses under “Learning Outcome Eight” (LO8: Reflect on and respond to community, national and global issues, as an informed and responsible citizen; Singapore Ministry of Education, 2014, 2016).

Inextricably linked to moral reasoning, the CCE curriculum intends for students to: (1) be able to distinguish right from wrong at the primary level, (2) exhibit moral integrity at the secondary level, and (3) have the moral courage to stand up for what is right at the pre-university level (Singapore Ministry of Education, 2014, 2016). The CCE syllabus also recommends that students at the secondary level should develop beyond the conventional toward the post-conventional level of moral reasoning according to the Stages of Moral Development (Kohlberg, 1984; see Table A1 in the Appendix), though it is noteworthy that his theory has not been without critics. For example, Kohlberg suggested that it was not always reasonable to retain the highest stage of the post-conventional level in his postulated model (i.e., Stage 6), and this caused some doubt in the applicability of his theory. However, it has been established that having few individuals exhibiting Stage 6 does not provide warrant for doubting the model. Further, some challenged the prescriptive and invariant nature of the stages within his model, though there has been significantly more studies.

1Singapore University of Social Sciences, Singapore
2Singapore Examinations and Assessment Board, Singapore
3The University of Western Australia, Perth, Australia

Corresponding Author:
Lyndon Lim, Teaching and Learning Centre, Singapore University of Social Sciences, 463 Clementi Road, Block C-9.04G, Singapore 599494, Singapore.
Email: lyndonlimjk@suss.edu.sg
supporting the notion that moral reasoning development tends to proceed invariantly (Lim & Chapman, 2021a).

Despite such syllabus shifts, it is unlikely that significant progress will be made in this area in the country as a whole in the absence of a valid means by which schools can monitor their progress toward specified goals. If schools are left to develop their own measures of moral development, it is likely that these measures will not have been validated. Such an approach, which would precipitate multiple non-standardized measures, would also not permit useful comparisons to be made across school contexts, or allow for data to be collected on a larger-scale basis. The latter goal must be reached for the Singapore MOE to determine the degree of progress being made in the country as a whole. There is, as a result, a need for an assessment instrument suitable for practical use in Singapore schools.

**Literature Review**

Evident through the works of many scholars including recent ones by Jacques and Anderson (2017), Choi et al. (2020), Coleman and Wilkins (2020), and Goldsmith et al. (2020), endeavors to assess moral reasoning in education (e.g., primary, secondary, and post-secondary educational institutions) have historically drawn upon established instruments such as the Moral Judgement Interview (MJI) by Kohlberg (1958), the Defining Issues Test (DIT) by Rest (1979) and its derivative DIT2 (Rest et al., 1999), the Ethical Reasoning Inventory (ERI) by Page and Bode (1980), and the Sociomoral Reflection Measure—Short Form (SRM) by Gibbs et al. (1982) and its derivatives (i.e., the Sociomoral Reflection Objective Measure [SROM] by Gibbs et al. (1984), the Sociomoral Reflection Objective Measure—Short Form [SRM-SF] by Basinger and Gibbs (1987), the Sociomoral Reflection Measure—Short Form [SRM-SFO] by Gibbs et al. (1992), and the Sociomoral Reflection Measure—Short Form Objective [SRM-SFO] by Brugman et al. (2007). A review of the literature found that all these instruments adopted either a production or recognition measure; a production measure requires respondents to construct a response either in text or verbally while a recognition measure requires respondents to select/rate/rank options. Empirical and secondary studies indicated that most of these instruments reported adequate levels of validity and reliability, with the exception of the SROM, SROM-SF, and SRM-SFO which found non-conclusive levels of validity and reliability in part due to the limited number of related studies.

While these instruments are clearly relevant for measuring moral reasoning, concerns such as content appropriateness and practicality do not favor their use in large-scale assessment programmes in the Singapore context (Lim & Chapman, 2021b). For instance, the MJI is an individually administered measure, and is thus not suited to broad-scale use in classroom settings, specifically for Singapore where class sizes commonly range between 35 and 42 students each. Other production measures, particularly those with less lengthy moral dilemmas such as the SRM-SF, are creditworthy for they afford some accessibility to students. However, the introduction of construct-irrelevant variance, a common threat to score validity, may not be ruled out given that not all students in Singapore may be fluent in English writing; most students are bi- or multi-lingual and might communicate using various ethnic languages with peers despite English as their lingua franca (Bolton et al., 2017). Further, production measures, such as the MJI or SRM, require scorers or teachers to undergo extensive training to minimize disparate scores (e.g., the MJI and SRM rater training require a minimum of 5 days and 6 hours, respectively), and these might present workload constraints for teachers, as opposed to recognition measures.

Recognition measures that are suited for such use from a practical standpoint, such as the DIT and its derivatives, are more fitting for the Singapore context though curriculum time of up to 50 minutes per class would have to be set aside if the DIT was administered. Further, they present scenarios (some of which have political content) and possible responses that may be difficult for non-native speakers of English from Singapore secondary schools to comprehend, in part due to the fact that students may not be able to relate to some of these hypothetical and complex scenarios. As an example, the SRM-SF presents scenarios related to contract, truth, affiliation, life, property, law, and legal justice; these may not be fully relatable to the intended audience of the MRQ. In fact, the developers of the SRM and its derivatives, Gibbs et al. (1992) conceded that the SROM-SF included fairly sophisticated moral dilemmas.

A few measures such as the SRM-SFO have found some preliminary success in reducing administration time (i.e., to 15 minutes). Nonetheless, more studies need to be undertaken before confirming the structure of the SRM-SFO given its more recent development and contrastingly fewer validity and reliability studies. Further, some of the established recognition measures that employed less confusing response formats have since been disused (e.g., ERI), judging from the corresponding dearth of literature. Other than the instruments discussed, there has been one developed by Soh (1987) for the Singapore context (i.e., Test of Moral Values). However, there has been a paucity of studies related to its validity evidence; more importantly, it was developed whilst not being anchored upon an established moral reasoning theory (e.g., Kohlberg’s stage-based theory) and hence, could not demonstrate any moral reasoning progression. Given these, it is not plausible to consider it for use.

In view of these, there is a clear need for the development of a validated, fit-for-purpose moral reasoning instrument that is suitable for use in Singapore school contexts. The following sections describe the procedures used to develop and validate the Moral Reasoning Questionnaire (MRQ), which
was designed specifically to assess Singapore secondary students’ (normally aged between 12 and 18) moral reasoning development based on Kohlberg’s stage-based model.

**Method**

The MRQ was validated in a series of three major stages, which corresponded with addressing the first three criteria stipulated by Messick (1993, 1995) in his elaboration of his well-established unified concept of validity. Messick stipulated that test developers, in evaluating the validity of their measures, should gather evidence related to: (i) test content; (ii) response processes required by the test; (iii) the internal structure of the test; (iv) associations between the test and other variables; and (v) the consequences of testing. Messick noted, however, that these were not to be interpreted as separate “types” of validity, just as different forms of evidence that can contribute to the overall evaluation of the validity of a measure.

Applying these notions to the current study, for the MRQ to be deemed an acceptable measure of moral reasoning, the scores generated by the MRQ should prompt meaningful interpretations about a respondent’s level of development in moral reasoning. The following sections present the validation of the MRQ in relation to evidence on the content of the MRQ (stage 1), on its response processes (stage 2), and on its internal structure (stage 3).

**Participants**

Three groups of participants were engaged for this study based on availability sampling. The first group participated in the initial content validation of the MRQ (stage 1 of the study) and comprised seven CCE experts who had experience within the Singapore education system ranging from 5 to 30 years. This group of experts participated voluntarily and signed the informed consent form, as required by the different institutional review boards of which the authors were affiliated with.

The second group (Group A of students) was involved in stage 2 of the study, and comprised five students selected to participate in a preliminary evaluation of the response processes associated with completing the MRQ. At the time of the study, one of these participants had completed secondary 2 and was 14 years old; one had completed secondary 3 and was 15 years old; and the remaining three had completed secondary 4 and were 16 years old. All five participants were from the express educational stream in Singapore, and attended different schools. Parental consent was sought for the five students in Group A to complete the 30 refined MRQ items by means of a participant consent form; each form was accompanied by an information sheet that specified that: (1) participant involvement in the research was voluntary, (2) participants were free to withdraw at any stage without prejudice in any way, with no reason required for withdrawal, and (3) all data would be anonymized and each participant would not be identifiable.

The third group of participants (Group B of students), was involved in Stage 3 of the study. This group comprised a large sample of secondary level students from Singapore schools. Access to the schools was granted by the MOE in September 2015. Of the six schools with different profiles invited via email to participate in this study, three agreed. As with the students in Group A, parental consent was sought before each Group B student attempted the MRQ by means of a participant consent form with the same participant information sheet used for Group A.

Across the three schools that eventually agreed to participate in this study (schools M, P, and X), the Group B sample comprised 670 Singapore secondary school students (497 females). The participants ranged in age from 12 to 18 years ($M=14.24$, $SD=1.30$ years) at the time of the study. Within the sample, 36.7% ($n=246$) of the students were drawn from school M; 44.2% ($n=296$) from school P; and the remaining 19.1% ($n=128$) were drawn from school X.

The three participating schools were diverse, representing different educational streams and educational levels. Of the 670 participants, 79.4% ($n=532$) were from the express stream, 16.3% ($n=109$) from the normal-academic stream, and 4.3% ($n=29$) from the normal-technical stream. In terms of education levels, 17.6% ($n=118$) of the participants were from secondary 1 at the time of the study; 27.8% ($n=186$) in secondary 2; 28.4% ($n=190$) in secondary 3; and 26.3% ($n=176$) were in secondary 4.

**Instrument Design**

**Item Stems**

Based on a thorough review of the literature on existing instruments assessing moral reasoning levels, a decision was made to retain the use of moral dilemmas as item stems (i.e., the “prompt” questions) within the MRQ. A large pool of items was generated initially, with the goal of providing meaningful differentiation across Kohlberg’s three levels of moral development, that is, pre-conventional, conventional, and post-conventional. The Kohlberg model, as noted previously, was adopted here in light of its strong alignment with the stipulated outcomes of the CCE syllabuses.

Christensen et al. (2014) noted that moral dilemmas have become a standard methodology for research on moral judgment. Evidently, most instruments to assess moral judgment or reasoning are based either partly or wholly on items that rely on moral dilemmas as prompts. One may argue that moral dilemmas are often unrealistic and do not indicate moral reasoning accurately. This can, however, be overcome by using more realistic and familiar dilemmas. In fact, in one early study, Sumfrer and Butter (1978) found no differences between college students’ levels of moral reasoning to hypothetical and actual moral dilemmas. This was also echoed in
a much more recent study by Bostyn et al. (2018), in which they concluded that responses to hypothetical dilemmas were predictive of affective and cognitive aspects of real-life decision making.

For the MRQ, moral dilemmas that incorporated prosocial, anti-social, and social pressure elements were used; these elements were drawn from a typology proffered by Krebs and Wark (1997), and considering what would be most applicable for the intended respondents. The moral dilemmas were developed with the goal of ensuring that they would be familiar and realistic to the intended respondents, to overcome problems associated with the “artificial” nature of the dilemmas.

Response Format

Ranking was selected as a response format as it requires respondents to order items based on some kind of identified attribute; in the case of the MRQ, this attribute would be moral reasoning. It has also been evident that scholars researching on values have consistently advocated the use of ranking for its suitability, based on the inherent comparative characteristic of values that allows respondents to demonstrate the degree of their views on an option relative to another (Alwin & Krohnick, 1985; Roccas et al., 2017). Further, Jacoby (2011) posed that ranking a set of values according to their subjective importance has “generally been regarded as the “gold standard” in obtaining empirical representations of individual value structures” (p. 2).

Nevertheless, there have been critiques on the use of ranking as a response format. For example, Clawson and Vinson (1978) proffered that, in using ranking as a response format, there was a possibility that “equally attractive values are forced into separate rankings” and “wide gaps in preference are treated as no different from minuscule gaps” (p. 398). However, based on Kohlberg’s stage-based model (where students progress based on mutually exclusive stages) and that the MRQ is not intended to form an evaluative judgement of the student’s moral reasoning level by focussing down to the decimal point of a score, the issues highlighted would not be critical. Further, Alwin and Krohnick (1985) stated that “rankings are often difficult and taxing for respondents, demanding considerable cognitive sophistication and concentration” (p. 536). Noting these critiques, care was taken to minimize sophistication and cognitive demands in the ranking options, owing to the intended age range of respondents for the MRQ. Items were worded with simple language and in short sentences in an attempt to minimize construct-irrelevant sources of variance in the responses.

The MRQ was designed primarily for delivery in online format, though an equivalent paper-based form was also developed and tested in the study. The examples presented in this paper are all taken from the online version of the instrument.

Based on a comprehensive review of the literature on response formats used in assessing moral reasoning, a decision was made to use a two-tier response format. Respondents first read the moral dilemma posed in each item. In Tier 1 of the response format, the respondent was required to select either of the “action” options posed (see Section 10a of Figure 1). The respondent was then presented with “moral reasoning” options on the subsequent screen depending on which option was selected in tier-one (see Sections 10b1 and 10b2 of Figure 1). In Sections 10b1 or 10b2 of Figure 1, the respondent was required to rank the options in order of importance to him/herself. Each of the options in 10b1 and 10b2 was designed to correspond with one of Kohlberg’s main levels of moral development.

Item Scoring

As the MRQ was designed to assess moral reasoning and not moral action, neither of the selected “action” options (i.e., Tier 1 responses) were scored. Scoring for Tier 2 responses is presented in Table 1.

Stage 1. Results and Discussion: Evidence Based on Item Content Appropriateness

Thirty possible items were prepared. Care was taken to avoid gender and race or religion bias when writing the items. Polysyllabic words were avoided in the moral dilemma vignettes. To the extent possible, sentences in the items averaged between 15 and 20 words, as recommended by Cutts (2013).

Over a period of 2 months, Group 1 participants were invited to critique and feedback on the content appropriateness of the items. In reviewing the items, the participants considered whether: (1) the content was suitable for 13 to 17 years old students with respect to moral reasoning, (2) the majority of 13 to 17 years old students would be able to understand the language structure and meaning of the moral dilemma and options in each item, (3) the options for each item were non-ambiguous and represented Kohlberg’s stages of moral reasoning, (4) the instructions on how to respond to the items were clear, and (5) whether the items were free of gender, race, or religion bias.

Items were subsequently refined based on the group’s feedback. For example, item 24 originally described a brother asking the respondent for money to buy textbooks. Of the panel experts posed that it would be unlikely that a brother would ask a secondary school student for money for a surgical procedure. Hence, this item was refined as a brother asking the respondent for money to buy textbooks.

Following the refinements and in agreement with the expert panel, all 30 items were deemed to have adequate content appropriateness.
Stage 2. Results and Discussion: Evidence Based on Response Processes

Upon receiving parental consent in Stage 2, each participant was interviewed individually by the first author throughout the process of completing the items. Notes were taken during this process so that these could be used in revising the item wordings. Widely used in the validation of questionnaire items (Beatty & Willis, 2007; Peterson et al., 2017; Willis, 2018), this form of validation (i.e., cognitive interviews) was used here to provide insight into the mental processes used by respondents in completing each MRQ item.

Each interview took approximately 1 hour and was recorded; participants were assured that all their responses would be anonymized and non-identifiable. Coupled with verbal probing, the processes used in these interviews followed the four-step model recommended by Ryan et al. (2012): (i) question and instruction comprehension; (ii) information retrieval (where clarity on how a participant recalls information to respond to the item is sought); (iii) judgment and estimation (where understanding on how participants arrive at a decision is sought); and (iv) documenting a response (whether a participant can match his/her response to the response-options provided by the items). Applying the model, participants first ranked the items as per the written instructions in the MRQ. They then justified verbally why

Figure 1. Example item with vignette and corresponding options.

Table 1. Scoring Matrix of Two-Tier Items.

| Rank order (levels of moral judgment) | Score |
|---------------------------------------|-------|
| (1) Post-conventional, (2) conventional, and (3) pre-conventional | 5 |
| (1) Conventional, (2) post-conventional, and (3) pre-conventional | 4 |
| (1) Post-conventional, (2) pre-conventional, and (3) conventional | 3 |
| (1) Conventional, (2) pre-conventional, and (3) post-conventional | 2 |
| (1) Pre-conventional, (2) post-conventional, and (3) conventional | 1 |
| (1) Pre-conventional, (2) conventional, and (3) post-conventional | 0 |
they chose those rankings and whether their responses reflected their thoughts.

The cognitive interviews revealed that the rank order of the options in Tier 2 were distinguishable by the respondents, and that the respondents did not have difficulties in ordering the options or choosing between the two “action” options presented in Tier 1 of each item. The interviewees also did not have any difficulty understanding the scenarios and no difficult words were highlighted.

Stage 3. Results and Discussion: Evidence Based on the Internal Structure of the Test

In Stage 3, 670 students from the three schools that agreed to support this study completed all 30 items in the MRQ. Students from two of the three schools completed the MRQ via the survey software platform Qualtrics™. Students from the remaining school completed an equivalent paper-based version of the MRQ due to logistical constraints identified by the school. Responses from the paper-based MRQ were input to the main dataset from Qualtrics™. All of the students whose parents gave consent completed the 30 items during curriculum time and within 40 minutes. Before the students did the questionnaire, teachers who helped to administer were given an information sheet and read a standardized set of instructions. Students were reminded that they could withdraw at any time and that their responses should be as honest as possible, as results of the MRQ would not affect their school results in any way.

For the purposes of the analysis, the total sample was split randomly into two halves, stratifying for school and class. Odd-numbered cases were used for all exploratory analyses performed, while even-numbered observations were used to confirm the results of the analyses performed on the odd-numbered cases. Each half of the sample was comparable in terms of gender, χ² (1) = 0.01, p > .05; secondary (grade) level, χ² (3) = 0.04, p > .05; age, χ² (6) = 6.15, p > .05; and academic stream, χ² (2) = 0.04, p > .05.

The analysis of the structure of the MRQ relied on two main procedures using SAS 9.4. Exploratory factor analysis (EFA) was first used to provide a preliminary assessment of the structure of the MRQ. Confirmatory factor analysis (CFA) was then used to determine whether the conclusions reached on the basis of the odd-numbered cases could be replicated with the even-numbered dataset.

Exploratory Factor Analysis

Prior to conducting the EFA, all assumptions underlying the use of this procedure were evaluated thoroughly to ensure that its use was tenable in this case. First, in terms of sample size, there were 335 odd-numbered observations available for use in the EFA and no missing data within the set. Recommendations for minimum sample sizes and/or case-to-variable ratios for EFA (e.g., Gable & Wolf, 1993; Goretzko et al., 2021; Mundfrom et al., 2005) vary so significantly within the education research literature that many scholars have argued that common “rules of thumb” for EFA sample sizes are not valid or useful (MacCallum et al., 1999; Watkins, 2018). In a synthesis of the literature related to minimum EFA sample sizes/case-to-variable ratios, Beavers et al. (2013) noted that the largest overall minimum sample size recommended in the papers they reviewed was n = 300, with a minimum case-to-variable ratio of 5:1. Given that both of these requirements were met within the current study, the use of EFA based on this dataset was deemed to be tenable.

Second, tests for univariate and multivariate normality on the MRQ total and item scores within the EFA dataset produced satisfactory results. For example, using the MRQ total scores, a very modest level of negative skew (−0.74) and a modest level of kurtosis (0.65) was obtained. Given that these coefficients fell well within accepted thresholds recommended by West et al. (1995), the assumption of normality was deemed to be tenable in this case. Similar coefficients were generated on the basis of the individual item score distributions within the EFA dataset. All other assumptions associated with the use of EFA, including linearity and the absence of outlying scores, were also evaluated, with results confirming that the use of EFA in this case was tenable.

A Maximum Likelihood (ML) estimation procedure was used as the primary factor extraction method within the EFA given its efficiency and ability in generating unbiased estimates, as suggested by some in situations of modest multivariate non-normality (Fabrigar et al., 1999; Sellbom & Tellegen, 2019). As noted previously, there were no significant violations of the normality assumption based on the preliminary investigations performed. However, to ensure that the very modest level of non-normality within the data distributions did not affect the factor extraction process, an Unweighted Least Squares (ULS) extraction (an extraction method that does not assume normal data distributions), was also performed for the purposes of cross-checking. As all of these analyses produced similar results to those obtained through the ML procedure, the paper focuses on results based on the ML extraction.

To determine whether the data was appropriate for the extraction of factors, Kaiser’s (1970) measure of sampling adequacy (MSA) was also computed. In this study, the MSA index was found to be 0.90 (for both ML and ULS methods), which is considered highly favorable (Hutcheson & Sofroniou, 1999; Kaiser, 1974). This confirmed that the use of EFA was appropriate in this case. Bartlett’s Test of Sphericity was also significant (χ² (435) = 2,315.37, p < .001), further confirming that the data were suitable for EFA.

Based on Kaiser’s (1958) eigenvalue rule, the initial analysis (without specifying the number of factors to be extracted) yielded a single strong factor that explained 76.72% of the
total score variance using ML estimation, and 75.88% using ULS estimation, for the entire set of variables. Visual inspection of the scree plots using either the ML or ULS estimation methods (see Figures 2 and 3) also provided warrant for the view that a single strong factor accounted for most of the variance in the MRQ scores.

To minimize the risk of under/over-estimation of factors and hence dimensionality in the interpretation of the MRQ results, a parallel analysis was then conducted with a Monte Carlo simulation of 1000 simulated datasets. Essentially, parallel analysis was developed based on Cattell’s scree plot to minimize issues related to factor indeterminacy (Bandalos & Finney, 2018; Franklin et al., 1995). While it is a less known procedure to determine dimensionality, it is recommended as the best procedure to assess the true number of factors as the Kaiser criterion may under/over-estimate the true number of factors (Basto & Pereira, 2012; Ledesma & Valero-Mora, 2007). The graphical output of the parallel analysis for this study is presented in Figure 4. The results of the EFA and parallel analysis were convergent in suggesting that the MRQ represents a unidimensional measure.

To explore specific items within the MRQ, loadings for each item were then examined, with a view to identifying any item with a loading less than 0.32 on the single factor.
retained, based on recommendations by Pasta and Suhr (2004), and Tabachnick and Fidell (2019). Based on the initial analysis, MRQ items A5, A11, A13, and A20 were removed, and a second EFA conducted using both estimation methods. Table 2 presents factor loadings from the second EFA which also yielded only one factor—which now accounted respectively for and increased 82.19% and 81.68% of the total score variance, using ML and ULS estimation methods. All factor loadings in this second EFA exceeded 0.32, with the minimum loading observed to be 0.35. Based on the EFA, it was concluded that the MRQ assesses one underlying construct, with 26 strongly loaded MRQ items.

Further evidence of internal structure of the MRQ was provided by the internal consistency data on the 26-item MRQ, using Cronbach’s $\alpha$ and item-total correlations. Results indicated a high level of internal consistency and minimal to no risk of item redundancy, based on recommendations by Hair et al. (2019) and Streiner (2003) (Cronbach’s $\alpha = .89$, item-total correlations ranging from .33 to .61). These results, together with those from the EFA, provided strong support for the unidimensional nature of the MRQ measure.

**Confirmatory Factor Analysis**

Confirmatory factor analysis (CFA) was used to ascertain the hypothetical model that the MRQ measured a unidimensional latent trait (moral reasoning). Preliminary analyses of the data indicated no missing cases, with univariate and multivariate tests for normality on the even-numbered observations (both for the MRQ total score and for the individual MRQ item scores) indicating skewness and kurtosis coefficients that fell well within thresholds recommended by West et al. (1995).

Various authors have noted that while CFA assumes multivariate normality, the ML method is generally robust to slight departures from normality (West et al., 1995). Nonetheless, the ML method may be unsuitable for severe violations of normality as the chi-square statistic may be over-estimated (Jackson et al., 2009). To reduce any risk of bias in the chi-square statistics obtained here, the ML with Satorra-Bentler scaled chi-square statistics for model fit (MLSB) was used in the study, as suggested by Motl et al. (2005).

For the purposes of further triangulation, comparisons of the results obtained with those using the ULS method were also conducted, given that ULS does not require any distributional assumptions (Jöreskog, 2003). In all, four CFAs were performed to ensure that the results were robust across a range of conditions. First, a CFA was performed using the MLSB method on the covariance matrix from the full original set of even-numbered data. Second, the MLSB method was used to perform a CFA on the MRQ, excluding items A5, A11, A13, and A20 (as suggested by the EFA results). Both these steps were then repeated using the ULS method, to confirm that the modest non-normality observed in the item distributions did not have any substantial impact on the results obtained. Table 3 presents the goodness-of-fit indicators for the four estimation methods.

The interpretation of these indices was guided by recommended thresholds of commonly reported goodness-of-fit indices by Hu and Bentler (1999), and those drawn from a review of CFA reporting practices by Schreiber et al. (2006), as well as recommendations by Hair et al. (2019) and Brown (2014).

Based on the absolute fit indices reported in Table 3, while the $\chi^2$ values obtained were significant, the $\chi^2/df$ values were less than 2 in both MLSB estimation cases. This indicated an acceptable overall model fit to the data despite the significant $\chi^2$ obtained, given the sensitivity of the $\chi^2$ statistic to sample size variations (Brown, 2014; Hair et al.,

---

**Table 2. Factor Loadings With Items Removed.**

| Item       | Loading on factor 1 | Item       | Loading on factor 1 |
|------------|---------------------|------------|---------------------|
| A27_score  | 0.65828             | A10_score  | 0.47641             |
| A14_score  | 0.61441             | A4_score   | 0.47540             |
| A21_score  | 0.57614             | A25_score  | 0.45869             |
| A26_score  | 0.56450             | A8_score   | 0.45352             |
| A24_score  | 0.55848             | A28_score  | 0.44563             |
| A18_score  | 0.55351             | A3_score   | 0.44528             |
| A16_score  | 0.52406             | A17_score  | 0.42846             |
| A30_score  | 0.52399             | A7_score   | 0.41866             |
| A12_score  | 0.51746             | A2_score   | 0.40766             |
| A23_score  | 0.50437             | A22_score  | 0.39791             |
| A19_score  | 0.49210             | A9_score   | 0.39659             |
| A15_score  | 0.47729             | A6_score   | 0.38961             |
| A29_score  | 0.47644             | A1_score   | 0.34928             |
The SRMR values of all MLSB and ULS estimation cases, which indicate the average discrepancy between correlations drawn from the data and those of the predicted model, were less than .05. These suggested a good model fit based on recommendations by Hu and Bentler (1999) and Hair et al. (2019).

The RMSEA was used here as the parsimony correction index, to provide an index of absolute fit of the model to the data (Hair et al., 2019). As the RMSEA values of both MLSB estimation cases were 0.04, falling between 0.03 and 0.04 with a 90% confidence limit, a good fit between the theoretical structure of MRQ and the empirical results obtained could be suggested.

Further, the CFI obtained for the MLSB estimation with items removed was 0.93, and was higher than that for the MLSB estimate with all items included. This suggested that, with items removed, the model fit was superior to that of the full item sample; the model fit with items removed was also relatively good. While the NFI is used less today, this index was also used here to provide information on the ULS estimation results. In this case, the NFI value both for the ULS estimation exceeded 0.96, which suggested a good model fit.

Further, the NFI for the ULS estimation with items removed was 0.97. This suggested again that a better model fit was achieved with these four items removed.

Collectively, the fit indices from the MLSB estimation with items A5, A11, A13, and A20 removed indicated a good overall model fit to the data. χ² (299, N=335) = 459.62, p < .001, SRMR = 0.05, RMSEA = 0.04, CFI = 0.93. Further, the slight non-normality of the data did not appear to have impacted the analysis outcomes, given that similar results were obtained through the use of the ULS estimation method (NFI = 0.97, SRMR = 0.05).

The standardized loading estimates on the latent trait (moral reasoning) and the standardized residuals of item pairs also supported the notion that the results indicated sound model fit to the data. All loadings obtained ranged from 0.42 to 0.66, and were statistically significant (p < .001), suggesting non-chance relationships between the items and the overall latent trait. Of the 26 items analyzed, three items (A2, A9, and A22) had loading estimates ranging 0.42 to 0.44; this range was slightly less than the 0.5 recommended threshold of standardized loading estimate by Hair et al. (2019). Standardized residuals provide information to identify item pairs for which the a priori measurement model does not accurately predict observed covariances between a pair of items. In this case, all standardized residuals of item pairs were less than 4.0, with the exception of the item pair A15 to A29 (4.49). These results would generally be considered acceptable, as recommended by Hair et al. (2019).

Further, the average variance extracted (AVE) and construct reliability (CR) coefficients were .3 and .9, respectively. While an AVE value of less than .5 suggests the possibility of more error variance in the items than variance explained by the latent factor on average, the CR coefficient exceeds the recommended .7. These suggest that the moral reasoning measurement model has adequate convergence and construct reliability (Hair et al., 2019).

### Limitations and Directions for Future Research

This study involved developing and validating the MRQ in a series of three major stages based on the unified concept of validity specified by Messick (1993, 1995). While the results based on EFA and CFA present the MRQ as fit for purpose, further studies should be undertaken particularly those associated with the fourth form of evidence supporting validity under Messick’s criteria (i.e., associations between the test and other variables).

To this end, the MRQ’s relations to other variables which were not evaluated in this study, given that the participating schools did not consent to having students undertake additional validation measures, should be reviewed subsequently. Future research, therefore, is needed to compare the MRQ with other variables. Two obvious choices would be the original Moral Judgement Interview (Kohlberg, 1984) and the Defining Issues Test-2 (Rest et al., 1999). These analyses would provide direct evidence on how the MRQ scores align with, or depart from, those obtained through other instruments designed to assess moral reasoning based on Kohlberg’s model. In addition to these measures, a range of other variables (e.g., cognitive ability, personality, and social attitudes) could also be used in the validation of the MRQ.

### Conclusion

In general, the evidence in this study provided both qualitative and quantitative support for the validity of the MRQ. In
light of these results, the MRQ could be applicable as a practical means to assess moral reasoning in Singapore secondary schools. As illustrated by Messick’s validity evidence criteria, it is recommended that the MRQ be used in conjunction with teachers’ classroom observation notes and anecdotes about students, as a means to validate the results of the MRQ. The MRQ may be applied beyond the context of Singapore, particularly if the context within which the MRQ is applied is not largely dissimilar to that of Singapore. Nonetheless, in relation to the content and evaluation of response processes of the MRQ based on Messick’s validity evidence criteria, studies could be undertaken to review the items and format under two suggested situations: (i) if the MRQ is used in a context vastly different from Singapore secondary classrooms; and (ii) whether information and communication technology could permit the items to be presented differently (e.g., presenting the moral dilemmas as cartoon strips instead of short paragraphs, to broaden the use of the instrument to other age groups, demographic groups, and/or contexts).

While further work can be done to provide further evidence on the generality of the MRQ, the validation analyses presented in this study suggest that this instrument has strong promise for use in the Singapore secondary context. Given the importance of moral education within the Singapore curriculum, it is imperative that schools have access to a validated, yet practical, means by which to assess the outcomes of schools’ efforts in this domain. The MRQ can serve such a purpose, which would not only allow individual schools to monitor their own progress in meeting the CCE LO8 goals, but also, permit valuable comparisons across schools and school contexts. The content of the MRQ is also quite general, and would be appropriate for use in other countries. It is anticipated, therefore, that the MRQ could have a significant impact in the area of moral education on an international basis.

Appendix

Table A1. Kohlberg’s Six Stages of Moral Development.

| Level          | Stage                              | Stipulated grade attainment |
|----------------|------------------------------------|-----------------------------|
| Pre-conventional | 1. Obedience and punishment orientation | K2–12                      |
|                | 2. Self-interest orientation        | K2–12                      |
| Conventional   | 3. Interpersonal accord and conformity | K4–12                      |
|                | 4. Authority and social-order maintaining orientation | K7–12                      |
| Post-conventional | 5. Social contract orientation     | K10–12                     |
|                | 6. Universal ethical principles    | K10–12                     |

Note. Adapted from http://www.psychologycharts.com/kohlberg-stages-of-moral-development.html.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval
Approvals to conduct this research were obtained from the Human Research Ethics Committee of the University of Western Australia (RA/4/1/7813) and from the Ministry of Education, Singapore (RQ105-15(09)).

Consent to Participate
Consent to participate was obtained from all participants either in hard-copy (in the case of qualitative interviews) or online (in the case of online instrument delivery) format.

Consent for Publication
Consent for publication was obtained in the same format as for consent to participate.

ORCID iDs
Lyndon Lim https://orcid.org/0000-0002-8199-5761
Elaine Chapman https://orcid.org/0000-0001-5861-1179

References
Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. Public Opinion Quarterly, 49(4), 535–552. https://doi.org/10.1086/268949
Anzai, S., & Matsuzawa, C. (2014). Values and value priorities underlying Japanese elementary-school moral education: Content analysis of Japanese elementary-school moral books. Mediterranean Journal of Social Sciences, 5(4), 359–369. https://doi.org/10.5901/mjss.2014.v5n4p359
Bandolou, D. L., & Finney, S. J. (2018). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), The reviewer’s guide to quantitative methods in the social sciences (pp. 93–114). Routledge. https://doi.org/10.4324/9781315755649

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%2Fper.1987.61.1.139

Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. Journal of Statistical Software, 46(4), 1–29. http://doi.org/10.18637/jss.v046.i04

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71(2), 287–311. https://doi.org/10.1093/poq/nfm006

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. Practical Assessment, Research & Evaluation, 18(6), 1–13. https://doi.org/10.7275/vq2q-rk76

Bleazby, J. (2020) Fostering moral understanding, moral inquiry & moral habits through philosophy in schools: A Deweyian analysis of Australia’s Ethical Understanding curriculum. Journal of Curriculum Studies, 52(1), 84–100. https://doi.org/10.1080/00220272.2019.1650116

Bolton, K., Botha, W., & Bacon-Shone, J. (2017). English-medium instruction in Singapore higher education: Policy, realities and challenges. Journal of Multilingual and Multicultural Development, 38(10), 913–930. https://doi.org/10.1080/01434632.2017.1304396

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleyes: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. Psychological Science, 29(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Brown, T. A. (2014). Confirmatory factor analysis for applied research. Guilford Press.

Brugman, D, Basinger, K. S., & Gibbs, J. C. (2007, August 11–14). Validation of the Sociomoral Reflection Objective Measure. In Advances in Consumer Research, 5(1), 396–402. https://www.acrwebsite.org/volumes/9454

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, J., Flexas, A., Calabrese, M., Gut, N., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. Frontiers in Psychology, 5, 1–18. https://doi.org/10.3389/fpsyg.2014.00607

Clawson, C., & Vinson, D. E. (1978). Human values: A historical and interdisciplinary analysis. Advances in Consumer Research, 5(1), 396–402. https://www.acrwebsite.org/volumes/9454

Coleman, R., & Wilkins, L. (2020). Moral development: A psychological approach to understanding moral decision-making. In W. Lee & G. C. Clifford (Eds.), The Routledge handbook of mass media ethics (pp. 43–72). Taylor & Francis. https://doi.org/10.4324/9781315545929

Cutts, M. (2013). Oxford guide to plain English. OUP Oxford. http://www.ebrary.com

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://psycnet.apa.org/doi/10.1037/1082-989X.4.3.272

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: A method for determining significant principal components. Journal of Vegetation Science, 6, 99–106. https://openaccess.library.calpoly.edu/cgviewcontent.cgi?article=1004&context=pb_pubs

Gable, R. K., & Wolf, M. B. (1993). Instrument development in the affective domain (2nd edn). Kluwer Academic Publishers.

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/10.1207/s1129963

Gibbs, J. C., Basinger, K. S., & Fuller, D. (1992). Moral maturity: Measuring the development of sociomoral reflection. Erlbaum.

Gibbs, J. C., Widaman, K. F., & Colby, A. (1982). Construction and validation of a simplified group-administrable equivalent of the Moral Judgment Interview. Child Development, 53(4), 895–910. https://doi.org/10.1207/s1129126

Goldsmith, J., Burton, E., Dueber, D. M., Goldstein, B., Sampson, S., & Toland, M. D. (2020). Assessing ethical thinking about AI. Proceedings of the AAAI Conference on Artificial Intelligence, 34(9), 13525–13528. https://doi.org/10.1609/aaai.v34i09.7075

Goretzko, D., Pham, T. T. H., & Büchner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. Current Psychology, 40, 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th edn). Pearson.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1–55. https://doi.org/10.1080/1070551990940118

Hutcheson, G., & Sofroniou, N. (1999). The multivariate social scientist: Introductory statistics using generalized linear models. SAGE Publications.

Jackson, D. L., Gillapsy, J. A., & Pure-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. Psychological Methods, 14(1), 6–23. https://doi.org/10.1037/a0014694

Jacoby, W. G. (2011). Measuring value choices: Are rank orders valid indicators? [Paper presentation]. Annual Meetings of the Midwest Political Science Association, Chicago, IL, United States of America. http://cces.gov.harvard.edu/files/cces/files/jacoby_measuring_value_choices-1.pdf

Jacques, K., & Anderson, S. (2017). Impact of a residence hall staff training class on the moral judgment development of college students. Open Access Library Journal, 4, 1–14. https://doi.org/10.4236/oalib.1104232
Singapore Ministry of Education. (2021). Desired outcomes of education. https://www.moe.gov.sg/education-in-sg/desired-outcomes
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
Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. Journal of Personality Assessment, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18
Sumfrer, G. F., & Butter, E. J. (1978). Moral reasoning in hypothetical and actual situations. Social Behavior & Personality: An International Journal, 6(2), 205–209. https://eric.ed.gov/?id=EJ215758
Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th edn). Pearson.
Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. Journal of Black Psychology, 44(3), 219–246. https://doi.org/10.1177/0095798418771807
West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues and applications (pp. 56–75). SAGE.
Willis, G. (2018). Cognitive interviewing in survey design: State of the science and future directions. In D. Vannette & J. Krosnick (Eds.), The Palgrave handbook of survey research. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-54395-6_14