Ethics Incognito: Detecting Ethically Relevant Courses Across Curricula in Higher Education

Martino Ongis[a], David Kidd[a], & Jess Miner[a]

[a]*Edmond & Lily Safra Center for Ethics, Harvard University*

Martino Ongis https://orcid.org/0000-0002-9410-1333
David Kidd https://orcid.org/0000-0001-6938-8291
Jess Miner https://orcid.org/0000-0002-2433-7893
All authors contributed equally to the article.
We have no conflicts of interest to disclose.
Link to OSF data: https://osf.io/25jvd/?view_only=f26d1844bc924fa08de18b7ebb1e45e7
Correspondence concerning this article should be addressed to Martino Ongis or David Kidd, Edmond & Lily Safra Center for Ethics, Harvard University, 124 Mt. Auburn Street, Suite 520-North, Cambridge, MA 02138, United States. Email: mongis@fas.harvard.edu or david_kidd@gse.harvard.edu

NOTE: This is not the final published version of the manuscript. To view the published manuscript in the *Journal of Academic Ethics*, see
https://link.springer.com/article/10.1007/s10805-023-09482-8

Ethics Incognito: Detecting Ethically Relevant Courses Across Curricula in Higher Education

As colleges and universities seek to invigorate ethics education, they need methods to identify where and describe how ethics is already present across their curricula. Meeting this need is complicated by the fact that much ethics education occurs in courses not explicitly focused on ethics or morality. In this paper, we review recent methodological advances before presenting a new Ethics Course Identification Tool (ECIT) that combines application of an expert-derived weighted dictionary and natural language processing methods to identify ethics-related courses based on their titles and course catalog descriptions, even when the terms "ethic" or "moral" are not present. Two studies, the second a pre-registered replication, revealed considerable interrater reliability among experts in ethics education regarding the ethical relevance of courses. Critically, both studies revealed strong correlations between expert judgments and ECIT scores. This empirical evidence points to a shared understanding of ethics education among experts, and it supports the valid use of the ECIT to rapidly and reliably identify ethics-related courses. Based on these findings, we propose that the ECIT can be used both to advance research on trends in ethics education and to help target interventions to improve ethics education at colleges and universities.

Keywords: ethics education; Word2Vec; expert judgment; interrater reliability; ethics across the curriculum; curriculum analysis

Introduction

Events and social changes of the past few decades have set off ethics education booms in disciplines not typically associated with the humanities, such as business, engineering, computer science, and biology (Elliott & June, 2018). Alongside the individual efforts of instructors to explore and respond to the ethical implications of their practice (e.g., Schouten, 2022), institutions of higher education at all levels, from individual colleges to inclusive associations and accreditors, have embraced their roles in supporting ethical development. The American Association of Colleges and Universities includes ethical reasoning as one of its essential learning outcomes for undergraduates (AAC&U, 2022), accreditors of all sorts require ethics instruction (Drechsler Sharp et al., 2011), and the mission, vision, and values statements of institutions nearly always express commitments to support moral and ethical development (Colby et al., 2003; Glanzer & Ream, 2008). Ethics training is also required for all researchers who work with human participants or non-human animal research subjects. Yet, empirical research reveals that students show little knowledge of, or concern for, ethics across different types of colleges and universities, prompting researchers to call for greater integration of ethics across curricula (e.g., Fischman & Gardner, 2022; Matchett, 2008).

What an ethics-infused curriculum might look like, though, remains unclear, as the definition of the term ethics itself is contested. We follow Elliott, who defines ethics as the discipline that concerns itself with how people act (or should act) in relation to subjects of moral worth (Elliott, 2007). Scholarly literature further tells us that ethics education broadly deals with helping learners clarify, prioritize, and integrate moral values (Weston, 2006), or pursue some conception of "the good" (Glanzer & Ream, 2008), which can be defined as a normative account of ideal behavior in a given sociocultural context. On the ground, however, definitions and details vary substantially across institutions (e.g., Colby et al., 2003), disciplines (e.g., Bebeau & Monson, 2008), and individual instructors (e.g., Kidd et al., 2020). Institutions with religious identities, for example, may be less likely to emphasize the ethical aspects of liberal participatory democracy than public institutions with missions rooted in serving local, state, or national communities, instead cultivating more universal ethics-related outcomes (Glanzer & Ream, 2008). Even within an institution, students in one course may find ethics narrowly construed in terms of compliance with professional codes of conduct, while students in another course may be encouraged to think critically about issues of social or environmental justice. Surely, ethics-related learning may be happening in both courses, but, as Matchett (2008) points out, encountering diverse perspectives on ethics may lead to relativism and cynicism about ethics if students fail to find coherence across their ethics-related learning experiences.

From this perspective, one of the key challenges to improving ethics education is to bring more intentionality and coherence to the ways ethics is addressed across the curriculum. To do this, it is critical first to identify where ethics-related teaching and learning is already happening. In this paper, we outline the challenges of identifying ethics-related courses and the limitations of extant methods before describing two studies designed to test the validity of an automated method for detecting ethics-related content across academic curricula.

Currently, most methods for studying ethics curricula require coding of courses based on the presence of a limited number of terms, often “ethic” and “moral”, in the course title, description, or syllabus, and these studies are usually focused on a specific discipline, such as journalism (e.g., Lambeth et al., 2004; Mills et al., 2019), engineering (Finelli et al., 2012), business (May et al., 2014; Wang & Calvano, 2015), or psychology (Griffith et al., 2014; You et al., 2018). More recently, Beever et al. (2021) extended these methods to an entire curriculum by applying an epidemiological approach to examine ethics education across a large, complex institution, although still defining ethics courses as those with “ethic” or “moral” in their titles or descriptions. Beever et al. (2021) show the value of taking a holistic view of ethics education across a campus, and their clear operational definition of ethics education promises transparent and reliable measurement.

In this case, however, reliability may come at the expense of validity because it hinges on a significantly constrained definition of ethics education. Research on student experiences and development consistently shows that ethics-related learning can happen anywhere students engage with ethical content, such as by studying diversity or practicing intergroup dialogue (for a review, see Mayhew et al., 2016), not just where ethics is the primary topic. Surveys of faculty and administrators bear this out, with many reporting embedding ethics across the curriculum (Christensen et al., 2007; Niell, 2017). As a consequence, searching only for courses that explicitly label ethics as a core topic will likely produce underestimates of the extent of ethics education; restricted studies of ethics education may generate distorted depictions of how institutions support ethical development.

Methods developed to operationalize a more expansive definition of ethics education face different limitations. For example, Bankhead et al. (2022) measured the presence of ethics in curricula by searching the course catalogs of 79 institutions for a set of 54 words related to ethics, with the presence of more words taken as an indicator of greater emphasis on ethics. Intriguingly, Bankead et al. (2022) report a positive relation between the frequency of these terms in course catalogs and students' performance on a test of moral reasoning, pointing to the importance of better understanding the presence of ethics across curricula. However, it is unclear if the analyses of word frequencies controlled for the total number of words in course catalogs, raising the possibility that the measure of moral language is confounded with the number of courses at an institution. Moreover, no evidence is offered to support the use of the particular 54 words as a measure of moral language, and it is unclear whether courses including these terms are consistently more related to ethics than courses without one of the 54 terms.

Using a much larger set of over 300 search terms, Kidd et al. (2020) developed a weighted dictionary search tool, the Ethics Course Identification Tool (ECIT), and presented evidence that instructors whose courses were identified as related to ethics tended to agree, providing initial support for the validity of the search method. Yet, the voluntary nature of the instructor survey makes it possible that instructors who disagreed with the search results simply neglected to participate in the study, possibly inflating the relation between the search results and instructors' perceptions. Moreover, as with the methods described by Bankhead et al. (2022), the original ECIT's validity hinges on two assumptions: First, that the expanded set of search terms adequately represents a wide range of approaches to ethics education, and, second, that there is a sufficiently shared understanding of ethics education to support reliable identification of courses that are not explicitly labeled as ethics courses.

The present research takes up the challenge of addressing these assumptions through the use of natural language processing methods to more fully account for the semantic associates of "ethics," and by directly testing whether expert raters agree with each other and the ECIT on the extent to which different courses are related to ethics.

Expanding the Ethics Course Identification Tool

To increase the sensitivity of the ECIT to a more expansive definition of ethics education, we analyzed the language used in a corpus of course titles and descriptions from a sample of 23 institutions within the Carnegie Classification system (the primary system for organizing U.S.

colleges and universities based on characteristics such as types of degrees offered and enrollment) indicating that they granted at least a bachelor's degree. The number of institutions selected from each classification was determined by the proportion of institutions (not students) in that classification within the population, though exact representativeness was not achieved due to difficulty identifying publicly available catalogs that could be converted into data files for analysis. This led to the random selection of 21 institutions, as well as two additional institutions that were added because their catalogs were already available to the researchers (see Table 1). This corpus included 62,880 course titles and descriptions.

We trained this corpus using Word2Vec, a technique for natural language processing that uses a neural network to learn word associations from a large corpus of text (Mikolov et al., 2013). This approach relies on the idea that related words appear close to one another in text passages. For example, the words "protects" and "safety" are likely to appear close to one another in texts because they are conceptually related – in other words, they appear in the same *semantic space*. Word2Vec estimates the prototypicality of target words based on their proximity to seed words in text passages and can be used to measure the similarity of target words like "security" and "shelter" to seed words like "protects" and "safety".

In addition to Word2Vec, we used distributed dictionary representations to identify the similarity between the concepts of ethics and morality and the course descriptions in our corpus. This method involves generating a continuous measure of similarity between a concept of interest, defined by a list of characteristic words, and any other piece of text (Garten et al., 2018). It has been used in psychological research on moral framing (Hoover et al., 2018) and the development of moral dictionaries (Hoover et al., 2020; Araque et al., 2020; for a review of some of these natural language processing approaches for psychological research, see Boyd & Schwartz, 2020). Our characteristic words, or seed words, were the words "ethics" (i.e., the E model) and "ethics" and "moral" together (i.e., the EM model).

This approach allows us to identify a large, representative sample of ethics-related courses quickly, and to identify ethical content even when ethics-related words are not explicitly used (such as in course titles and descriptions that do not use any of the terms included in the expert-driven ECIT). The advantage of this approach is that Word2Vec scores are not based on predetermined expert criteria, but rather on linguistic patterns identified in the original corpus of course titles and descriptions, potentially enabling a more comprehensive and unbiased identification of ethics-related courses. Moreover, we can use this method to give each course in our sample a score indicating how closely it is associated with the concepts of ethics and morality.

Research Overview

Can our data-driven tool accurately identify ethics-related content in course catalogs? And do human judges agree on what constitutes ethics-related content in a curriculum? In a previous study, Kidd et al. (2020) found that the dictionary-based, expert-driven ECIT was effective at identifying ethics-related content. However, we wanted to know if the machine-learning, data-driven ethics would also be able to identify such content, and if it could overcome the limitations of the expert-driven ECIT. Additionally, we were curious if there was a "shared perception" of ethics-related content among human judges, and if the ethics-related content identified by the ECIT aligned with this perception.

We conducted two studies to determine whether the machine-learning, data-driven ECIT could accurately identify ethics-related content in course catalogs. We also wanted to determine whether human judges would agree on what constituted ethics-related content, and whether the ethics-related content identified by expert-driven or data-driven ECIT aligned with the perceptions of such content held by human judges. First, we tested whether experts in ethics education agreed on what constituted an ethics-related course based on limited information (i.e., course title and description). This allowed us to determine the extent to which the average of their ratings reflected a shared understanding of ethics-related content. Second, we tested the validity of the data-driven ECIT by comparing the courses identified as ethics-related by the tool with the perception of such courses by experts in ethics education. Specifically, we asked these experts to rate the extent to which they believed a set of randomly selected course titles and descriptions were ethics-related, and then compared these ratings to the ECIT scores for the same courses.

In Study 1, we aimed to determine the extent to which experts in ethics education agreed on the content of ethics-related courses based on limited information (course titles and descriptions), and we sought to test whether these human held perceptions would correlate with automatic scores on the ECIT. To do this, we randomly selected a number of course titles and descriptions from the larger corpus we used to develop the data-driven ECIT, and we asked participants – all of whom were familiar with ethics education – to estimate the extent to which they thought a particular course was ethics-related. In Study 2, we conducted a pre-registered replication of Study 1 with a more focused sample of experts on ethics education. We asked each participant to rate a greater number of courses than the ones used in Study 1, and we further investigated the consistency and the shared understanding of ethics-related content among experts in the field, as well as the relationship between human perceptions and automatic scores on the ECIT. For both studies, analyses were conducted only after data collection was complete. The research reported in the paper was conducted in compliance with the field's ethical standards and was approved by the Institutional Review Boards of the authors' institutions. All of the materials and data can be accessed through the Open Science Framework: https://osf.io/25jvd/?view_only=f26d1844bc924fa08de18b7ebb1e45e7

Study 1

Methods

ECIT selection criteria. The corpus of course titles and descriptions we used to select courses for the rating task was the same used to develop the data-driven ECIT. To test the relationship between ECIT scores and expert ratings, we developed a set of criteria that helped us to randomly select courses from our corpus with an equal number of courses with high or low scores on the expert-driven and data-driven ECIT. We therefore assigned each course in our corpus with a label (either "high" or "low") depending on the following criteria: (i) a course scoring 1 or above on the expert-driven ECIT was labeled as "expert-driven high;" and a course scoring 0 on the expert-driven ECIT was labeled as "expert-driven low;" (for more details on the scoring system of the expert-driven ECIT, see Kidd et al., 2020) (ii) a course scoring in the top 10% of the data-driven ECIT was labeled as "data-driven high;" a course scoring in the bottom 50% of the data-driven ECIT was labeled as "data-driven low." We randomly selected an equal number of courses (n = 18) from each of the resulting four brackets.

The category high-high (high scores on both the data-driven and expert-driven ECIT) consisted of a pool of 2,524 courses, of which 1,849 contained either the words "ethic*" or "moral." Therefore, this category included 675 course descriptions that would have not been identified with only these keywords. Among the 18 courses that were randomly selected for the study in this category, the words "ethic*" and "moral" appeared in the title or course description of 9 courses. In other words, the titles and descriptions of half of the courses in the high-high category did not explicitly include the terms "ethic*" and "moral." An example of a course that was selected for the study that did not explicitly include these terms was a course titled "Topics in Ancient Philosophy: Plato and Aristotle on Art and Rhetoric," with the following description: "Plato's and Aristotle's views on the nature of art and rhetoric and their connections with the emotions, reason and the good life. Readings include Plato's *Gorgias*, *Ion* and parts of the *Republic* and the *Laws* and Aristotle's *Poetics* and *Rhetoric*." The category low-low (low scores on both ECITs) consisted of a pool of 26,007 courses, none of which included in the title or description the terms "ethic*" and "moral." The category high-low (high scores on the data-driven ECIT, low on the expert-driven ECIT) included 2,245 courses, none of which had the terms "ethic*" and "moral" in the title or description, whereas the category low-high (low scores on the data-driven ECIT, high scores on the expert-driven ECIT) had 753 courses, 19% (n = 145) of which contained the explicit terms "ethic*" and "moral" in their title or description.

Participants and procedure. Participants (n = 61) were recruited via email using the contact lists of the National Ethics Project (https://nationalethicsproject.org/) and the Association for Practical and Professional Ethics (APPE; https://www.appe-ethics.org/), and were asked to voluntarily take part in a research study to be taken online on the survey platform Qualtrics. Each participant was randomly assigned to one of six blocks. Each block included course descriptions randomly selected from four sets of course descriptions, reflecting high and low scores on the expert-driven ECIT approach and high and low scores on the data-driven ECIT approach. For each block, three course descriptions from each of these four sets were randomly selected, yielding 12 course descriptions in each block. For each course title and description, participants were asked to rate the extent to which they thought it was related to ethics on a five-point scale (1-*Not at all related to ethics*; 2-*Barely related to ethics*; 3-*Somewhat related to ethics*; 4-*Considerably related to ethics*; 5-*Absolutely related to ethics.*) Before rating the courses, participants were asked to report some demographics information (i.e., age, gender, race/ethnicity), their current position (e.g., tenure or tenure-track professor, graduate student, etc.), their level of education, their current academic institution type (e.g., public, private, etc.), their type of involvement in ethics and ethics education (e.g., designed or taught an ethics course, contributed to literature on ethics education), the extent to which they considered themselves experts on ethics education on a 5-point Likert scale ranging from *"1-Not at all"* to "*5-To a very great extent,"* and their discipline(s) of expertise. Moreover, they were asked whether they were members of APPE.

Given the potential for incomplete responses or for the need to remove non-experts from the pool of participants, we attempted to obtain at least 10 participants for each set of 12 courses. To achieve this, we began the study with only two blocks of 12 course descriptions, to which participants were randomly assigned. Once we observed at least 20 responses, we planned to replace the first two blocks with the second two, and so forth. However, we were unable to intervene exactly at the point 20 responses were obtained, leading to a greater number of responses for the first two blocks than for the others (see Table 2). Overall, we obtained ratings for 72 unique courses, albeit some of these courses had more ratings than others.

Results

Inter-rater reliability. To address our research question regarding the relation between expert ratings of the relevance of different courses to ethics education, we needed to evaluate the consistency of expert ratings. To evaluate inter-rater reliability, we calculated the intraclass correlation coefficient (ICCs) of ratings of 12 courses. We calculated two different ICC values to reflect the two purposes of evaluating interrater reliability. Our first goal was to estimate the extent to which experts agree on what constitutes an ethics-related course. To generalize to other similar raters, we adopted a two-way random effects model. Since we wanted to know the extent to which any single expert on ethics education would agree with others, we specified a single rater, and we specified that we are primarily interested in the consistency of the relative ratings of course descriptions, rather than absolute agreement on the precise value of the rating. Using Koo and Li's (2016) recommended method of reporting ICC, we calculated ICC estimates based on a single rating ($k = 1$), consistency, 2-way random-effects model. All analyses were conducted using the Intracc Macro for SAS (Hamer, 1990).

Our other purpose is to evaluate the reliability of the average ratings given the number of raters for each block. Insofar as these average ratings are reliable, they can be used to test whether the automated search methods provide a valid measure of the relevance of courses to ethics education. To assess the reliability of the average scores, ICC estimates were calculated based on a mean rating ($k = n$), consistency, 2-way mixed-effects model.

Overall, the average observed ICC or a single rating ($k = 1$), consistency, 2-way random-effects model was .59, which is close to the conventional standard for good reliability (Cicchetti, 1994), though there was substantial variability across blocks, with Block 3 having an especially low ICC (see Table 2). Considering that participants were given minimal instructions and no training, these findings suggest that experts in ethics education tend to agree on the likelihood that a course is relevant to ethics education based on its catalog description.

The lowest ICC was observed in Block 3, which had 11 raters. A post-hoc power calculation (Zou, 2012) indicated that 90% power to distinguish this ICC from zero would be obtained with only seven ratings of courses, making the ICC analysis adequately powered, despite the low observed ICC. Power can also be influenced by the number of raters, and the smallest number of raters was 4, in both Block 5 and Block 6. Of these two blocks, Block 6 was associated with the lowest ICC. Again, however, a post-hoc power analysis confirmed that an ICC of .54 or higher could be distinguished from zero with 90% power with 11 course ratings. In sum, despite variation in observed ICC values and different numbers of raters, the six ICC analyses all had sufficient statistical power to yield reliable estimates.

Correspondence of expert ratings and ECIT scores. First, we tested whether expert participants gave different ratings to courses selected from different sets of course descriptions. Specifically, we expected that course descriptions with high scores on both the original ECIT and the new search tool would be rated as most likely to address ethics, followed by course descriptions with a high score on either of the search tools. We expected courses selected from the set of courses receiving low scores from both search tools would be rated as the least related to ethics. A within-subjects ANOVA was conducted in which the set from which courses were selected was entered as a within-subjects independent variable and course rating was entered as the dependent variable. There was a significant effect of course set ($F(3, 180) = 126.18$, $p < .0001$), and follow-up within-subjects analyses revealed that all sets differed significantly from each other (Fs > 30.66, ps < .0001). As expected, courses selected from the set of course descriptions receiving high scores from both search tools were rated as most related to ethics (see Table 3 for means and standard deviations).

Second, we tested the relationship between expert ratings and ECIT scores using course descriptions as units of analysis ($n = 72$). We expected expert ratings to correlate positively with both expert-driven and data-driven ECIT scores, and as expected, we found a positive correlation between expert ratings and expert-driven ECIT scores ($r(72) = .61$, $p < .0001$) and between expert ratings and data-driven ECIT scores ($r(72) = .49$, $p < .0001$). To test the unique contribution of each component of the ECIT, we ran a general linear model predicting expert ratings using both ECIT scores as predictors. The overall model was significant, $F(2, 69)$ =

48.18, $p < .0001$), but more importantly, both the expert-driven ($t = 5.71$, $p < .0001$) and the data-driven ECIT ($t = 5.82$, $p < .0001$) predicted a unique portion of the variance in expert ratings. Despite their conceptual similarity, both components of the ECIT picked up something unique about ethics relatedness.

Exploratory Analyses Using Two Seed Terms for the data-driven Dictionary. Although using the seed term "ethic" appears to be the most simple and direct way of defining the semantic space of ethics in course catalogs (the E model), we also tested the model we created using both "ethic," and "moral" as seed terms (the EM model) to see if it would perform better than the E model. Scores derived from this EM Model were more strongly correlated with expert ratings ($r(72) = .618$, $p < .001$) than were the scores derived from the E Model ($r(72) = .490$, $p < .001$), suggesting the more inclusive model may improve the performance of the ECIT. For example, the following course received a low score using the E Model but high scores from the expert-driven method and participants in Study 1:

> The Good
>
> We will consider some recent and historical work on the Good, in order to answer such questions as: what is the relation between something's being good and something's being good for someone; whether what is good for someone is relative to his nature; whether we always act "under the guise of the good;" whether goods can be aggregated across the boundaries between individuals; what are the criteria by which final ends and lives may be judged good; what kinds of things (people, animals, plants, nation-states, ecosystems, species?) have a good that matters morally; what is the relation between being morally good and having a good life; and of course, what is the Good?

Using the EM Model, this course receives a much higher score (0.318) than it receives using the E Model (0.160), bringing the data-driven method into stronger agreement with expert participants. Accordingly, the EM Model was used to select courses in the design of Study 2.

Discussion

Results from Study 1 provide initial support for all hypotheses: experts in ethics education tended to agree with each other and with the ECIT on the relevance of courses to ethics. However, several limitations informed the design of Study 2. First, exploratory analyses suggested superiority of the EM Model, which was selected to replace the E Model in Study 2. Second, Study 1 included a small number ($n = 72$) of courses as stimuli, raising questions about the generalizability of the findings. Given the small number of raters needed to obtain reliable expert ratings, a more limited group of participants was asked to rate a much larger ($n = 128$) set of courses. Third, the broad call for anonymous participation by individuals interested in ethics education means that there is no way to guarantee that participants truly hold expertise in the area. In Study 2, personalized invitations were sent to a small group of scholars within the research team's extended professional network. These scholars were defined as having expertise based on their leadership of ethics centers, ethics associations, or ethics education interventions across the United States. Although the more limited sample may compromise representativeness of the larger population, it helps bolster the internal validity of Study 2 by ensuring high quality responses from recognized experts. Finally, the necessarily exploratory nature of Study 1 makes it critical to conduct confirmatory studies closely replicating its key findings (Nosek & Linday, 2018; Van't Veer & Giner-Sorolla, 2016).

Study 2

Methods

Pre-registration. For Study 2, we pre-registered five hypotheses based on the pattern of results we observed in Study 1. We hypothesized that participants would demonstrate acceptable interrater reliability (i.e., ICC > .60) when rating courses (Hypothesis 1). Next, we hypothesized that average ratings of courses would positively correlate with the expert-driven ECIT and with the data-driven ECIT (Hypothesis 2). We further hypothesized that when expert ratings are regressed on the two search tools, both tools would account for significant variation (Hypothesis 3). Next, we hypothesized that ratings of courses with high scores on both search tools would be significantly higher than those with a high score on only one tool or low scores on both tools (Hypothesis 4). Finally, we hypothesized that ratings of courses with a high score on at least one search tool would be significantly higher than ratings of courses with low scores on both search tools (Hypothesis 5). A pre-registration of the study's methods, analyses, and exclusion criteria can be found at: https://aspredicted.org/k7f49.pdf

Participants and procedure. Participants were recruited via personalized emails and were selected through personal contacts by one of the authors (J.M.) based on their level of expertise on ethics education. None of the participants for Study 2 had participated in Study 1. Based on evidence gathered in Study 1, we determined that three raters yield acceptable ICC for average ratings. A-priori power analyses indicated that 95% power to replicate observed effect sizes in our initial study for Hypothesis 2 and Hypothesis 3 would be obtained with at least 37 courses.

Accordingly, we recruited three participants for each of four blocks of courses. The total number of participants we recruited was 12, and the total number of rated courses 128.

Of the 32 courses in each block, 8 were randomly selected from the bracket with high scores on both ECIT tools, 8 from the bracket with low scores on both ECIT tools, 8 from the bracket with high scores on the expert-driven ECIT and low scores on the data-driven ECIT, and 8 from the bracket with high scores on the data-driven ECIT and low scores on the expert-driven ECIT. The procedure for selecting courses with high scores on the data-driven ECIT was the same as in Study 1, except that the EM Model was used instead of the E Model.

As in the first study, participants completed the study online via Qualtrics, where after responding to the same set of questions used in Study 1 they were randomly assigned one of four blocks, each containing 32 course descriptions. For each course title and description, participants were asked to indicate the extent to which they thought each course was related to ethics on the same Likert-scale used in Study 1, ranging from "*Not at all related to ethics"* to *"Absolutely related to ethics*".

Results

Interrater reliability. Goals were the same as the ones outlined in Study 1: estimate the extent to which experts agree on what constitutes an ethics-related course and evaluate the reliability of the average ratings given the number of raters for each block. As in Study 1, we calculated ICC using Koo & Li's (2016) method, and we conducted the analyses using the Intracc Macro for SAS (Hamer, 1990).

Overall, the average observed ICC or a single rating (k = 1), consistency, 2-way random-effects model was .67, with some degree of variability across blocks (see Table 2). These findings confirmed our first hypothesis: ICC was well above the acceptable threshold (.60), meaning that experts in ethics education tend to agree on the likelihood that a course is relevant to ethics education based on its catalog title and description. The lowest ICC was observed in Block 1 (.60) whereas the highest ICC was observed in Block 3 (.71). Our analyses had sufficient statistical power to yield reliable results.

Correspondence of expert ratings and ECIT scores. As in Study 1, we expected that course descriptions with high scores on both the original ECIT and the new search tool would be rated as most likely to address ethics (Hypothesis 4), followed by course descriptions with a high score on either of the search tools (Hypothesis 5). Courses receiving low scores from both search tools would be rated as the least related to ethics. First, We conducted a within-subjects ANOVA in which the within-subjects independent variable was the set from which courses were selected and the dependent variable was course rating. There was a significant effect of course set ($F(3, 33) = 88.88$, $p < .0001$). Follow-up within-subjects analyses were run to test our predictions. As in Study 1, all sets differed significantly from each other (Fs > 42.31, ps < .0001). As expected, courses selected from the set of course descriptions receiving high scores from both ECIT components were rated as most related to ethics, followed by courses that scored high on either one of the ECIT components and, finally, courses that scored low on both ECIT components (see Figure 1).

Correspondence of expert ratings and ECIT scores. We hypothesized that average ratings of courses would positively correlate with both the expert-driven and the data-driven ECIT tool, as observed in Study 1 (Hypothesis 2). As predicted, we found a positive correlation between expert ratings and expert-driven ECIT scores, $r(128) = .67$, $p < .0001$, and between expert ratings and data-driven ECIT scores, $r(128) = .50$, $p < .0001$. Unlike in Study 1, the relationship between the expert-driven and the data-driven ECIT scores was only *marginally* significant, $r(128) = .14$, $p < .10$. We also predicted that both the expert-driven and data-driven ECIT scores would account for an independent portion of the variance in predicting expert ratings (Hypothesis 3). A multiple regression analysis revealed that this was indeed the case, $F(2, 125) = 100.86$, $p < .0001$; $t_{\text{expert-driven}} = 10.92$, $p_{\text{expert-driven}} < .0001$; $t_{\text{data-driven}} = 7.39$, $p_{data\text{-}driven} < .0001$. Thus, as expected, the expert-driven ECIT and the data-driven ECIT picked up an independent component of ethics-relatedness.

Discussion

In Study 2, we conducted a pre-registered replication of Study 1 with a more focused sample of participants and a greater number of courses. Our five hypotheses were all confirmed in the results. We found that participants demonstrated acceptable inter-rater reliability and that the average ratings of courses by human experts were correlated with both components of the ECIT. In addition, both components of the ECIT were found to be significant predictors of expert ratings, and courses that received high scores on both components were rated as the most related to ethics, followed by courses that received a high score on only one component. These results replicate and extend the findings of Study 1, providing further evidence for the validity and utility of the ECIT as a tool for identifying ethics-related courses.

General Discussion

The results of the second study offer strong support for the hypotheses derived from the findings of the first study. The cumulative evidence appears to confirm the hypothesis that experts in ethics education broadly agree on the relevance of different courses to ethics, even when ethics and morality are not explicitly mentioned in the course title or description. Although the single-rater reliability only met the minimum standards for social science research, it is important to highlight that participants received no training or guidance beyond the brief instructions in the survey. The results do not suggest consensus, but they do show enough consistency across judges to infer a shared definition of ethics education within the academic community, even in the absence of a clear definition. Put simply, if one expert thinks a course is relevant to ethics, one of their colleagues is likely to agree.

Drawing on Pike's (2015) conceptions of *etic* and *emic* descriptive approaches, this finding suggest that, although there is no clear etic, or universal expert-driven, definition of ethics education, there may be a more implicit emic understanding of the meaning of ethics shared within the community of scholars interested in ethics education. In this context, Berry's (1989) application of Pike's (1954) concepts to efforts to describe intelligence in cross-cultural psychology provides a useful example. Historically, researchers attempting to study intelligence across cultures typically started with seemingly etic, or *pseudo*etic (Triandis et al., 1973), theories of the construct that to them appeared universally applicable but failed to adequately represent how people in other cultures viewed intelligence. An emic approach to defining intelligence, though, would start with examining how people in different communities understand intelligence, using their judgments as a basis for evaluating the validity of any measure of intelligence. This method has been used in studies of language fluency (Barnwell, 1989) and lexical diversity (Jarvis, 2017); the present results suggest that it can be productively extended to research on ethics education, which, like intelligence or fluency, is a socially defined construct.

Consistent support for the remaining four hypotheses shows that expert judgments on ethics education can be substantially reproduced using a combination of two methods for text analysis, expert-driven dictionary creation and a data-driven, natural language processing approach to building a dictionary. The studies were designed to test the unique contributions of each method by treating each as a separate, two-level factor. Unsurprisingly, when the two search methods agreed, experts were most likely to agree. When only one search method identified a course as ethics-related, experts were still likely to agree, but with less clarity. This finding highlights the value of combining expert-driven with data-driven methods for dictionary creation (Garten et al., 2018), since exclusive use of either method would result in a higher number of false negatives.

One set of courses received high scores from the data-driven dictionary but low scores using the original expert-driven dictionary. As expected, application of the data-driven dictionary helped identify courses with very short titles and descriptions, addressing a key limitation of the original, expert-driven dictionary. It is important to note, though, that this was a very large category of courses, many of which received relatively low ratings from experts. To some extent, low ratings may stem from a lack of information, but it may also be that some courses address ethically relevant issues, such as counseling or policing, but fail to engage with their ethical aspects, even subtly, in their titles and descriptions. In future research, adding questions about whether a given course should address ethics, rather than just whether it appears to address ethics, may clarify the extent to which these courses are indeed ethically relevant.

Another set of courses met the threshold of the original expert-driven dictionary but received relatively low scores on the new data-driven dictionary. This smaller set of courses tended to receive high ratings from experts, who likely inferred ethical relevance due to the presence of key terms identified by other experts as related to ethics (see Kidd et al., 2020). The high expert ratings are interesting because these course descriptions and titles include relatively little language that often appears in close proximity to “ethic” (and, in Study 2, “moral”) across the large corpus of course catalogs used to develop the data-driven dictionary. Experts readily identify these courses as related to ethics, but their titles and descriptions evade detection using the Word2Vec model, which may primarily reflect traditional and concretized approaches to ethics education.

Expanding the semantic space reflected in the data-driven method by using both “ethic” and “moral” as seed terms, as was done in Study 2, does not eliminate cases in which the original expert-driven method and the experts recognize a course as ethically relevant despite it receiving a low score using the data-driven method. A review of titles from this set of courses suggests they tend to address emerging topics in ethics education, such as social justice, the environment, and new technologies, compared to the more traditional courses focused on philosophical or professional aspects of ethics identified by both dictionary-based search methods (and experts) as related to ethics. Future research could more formally explore a larger sample of courses from this category to identify, describe, and track the spread of new topics in ethics education. If these topics become more common in ethics education over time, they should eventually be more prominent in the semantic space of "ethic," and "moral," and courses addressing them would then be identified also using a dictionary derived from a data-driven method.

At the same time, updates to the expert-driven dictionary can be made to identify new emerging topics or improve construct coverage. Initially, users of the ECIT may supplement the tool by manually searching for specific terms they believe should be added to the expert-driven dictionary. For example, a reviewer of this manuscript noted that including the phrase "service learning" in the expert-driven dictionary would be helpful because service learning courses are often intended to support ethical development (e.g., Sahatjian et al., 2022). Although the process for developing the expert-driven dictionary (see Kidd et al., 2020) did not lead to the inclusion of this term, a researcher could independently search for this term to ensure they identify course titles and descriptions including it. As other researchers use the ECIT and propose additions, the expert-driven dictionary could be expanded to formally include new terms that are commonly added by researchers. After such revisions, though, additional evidence would be needed to evaluate the validity of the updated tool. Accordingly, we recommend that ad hoc additions to the ECIT be made explicit to ensure its valid use and support its further development.

Limitations

A key aim of these studies was to evaluate the extent to which members of a broadly defined community of ethics educators share an understanding of what constitutes ethics-related teaching and learning despite lacking a consensus-based explicit definition. However, the recruitment methods and nature of the study may have led to a sample of ethics educators open to an expansive view of ethics education. Ethics educators who interpret the domain of ethics education narrowly in terms of prescriptive codes of conduct, for example, may not have been reached through our recruitment efforts or may have been less interested in participating. Moreover, despite efforts to recruit ethics educators from a range of institutions and disciplines, the relatively small sample sizes require some caution when generalizing to the entire population. A similar limitation applies to attempts to generalize these findings to the full population of course descriptions. Accordingly, the present evidence, however encouraging, needs to be corroborated by future studies.

To be used with validity, the ECIT's limitations also need to be kept in mind. The search method, though efficient and reliable, is not intended to be used to definitively determine whether or not a course is relevant to ethics education, much less whether it actually contributes to learners' ethical development. The ECIT is also unable to characterize the specific learning objectives or the depth of ethical inquiry in a course. Rather, the ECIT is intended to identify courses in a curriculum that relevant experts would likely agree is probably related to ethics. Further inquiry, using either surveys (e.g., Kidd et al., 2020), interviews, or analysis of artifacts (e.g., syllabi and student work) would be needed to more fully understand a course's contributions to ethics education.

Another limitation of the ECIT is that it, and its supporting evidence, all come from institutions of higher education within the United States. Variation in approaches to education and, of course, linguistic differences would likely make using the ECIT inappropriate in other contexts. To some extent, these concerns are relevant even within the United States, especially given the limited number of course catalogs constituting the corpus used to develop the ECIT. The sample may not adequately represent approaches to ethics important in more specialized schools (e.g., tribal colleges and universities, military academies), and future iterations of the ECIT could be improved by basing them on a more expansive corpus. Some of the patterns around ethics identified by the underlying model used to develop the ECIT may not have direct equivalents in other languages, and direct translation may introduce biases. We therefore encourage researchers interested in other linguistic or cultural contexts to replicate the methods for developing the ECIT, rather than translating the English-language ECIT.

Intended Uses

As a research tool, the ECIT can be used to facilitate various lines of inquiry related to ethics education. Researchers can use the ECIT to explore the prevalence and characteristics of ethics-related courses across different disciplines and institutions. For instance, the ECIT can be used to compare the number and distribution of ethics-related courses offered at liberal arts colleges versus research universities, or examine how the content and pedagogy of ethics courses vary across different departments. In addition, the ECIT can help researchers identify sites for assessing the impact of ethics education on learners' moral reasoning and ethical development.

By identifying which courses are more likely to address ethical issues and using measures of ethical development before and after a particular course is taken (for some of these measures, see Kidd et al., 2020), researchers could measure the effectiveness of ethics education in promoting moral reasoning and ethical development.

Beyond its uses as a research tool, the ECIT may also have practical applications in higher education. It can be used to identify where ethics is being addressed, allowing faculty and administrators to evaluate the coherence and breadth of ethics education across a campus, identify gaps in the curriculum, and develop targeted strategies to improve the quality and impact of ethics education. For instance, administrators and researchers can assess the robustness of the existing offerings of ethics-related courses across their academic catalogs and more quickly implement new programs, such as an Ethics Across the Curriculum program. This application could be particularly valuable for institutions working to meet accreditation standards related to ethics education, as the ECIT could provide a systematic and comprehensive way to demonstrate compliance and continuous improvement. Of course, one should not forget about the limitations of the current version of the ECIT when using it for both research and practical purposes.

To facilitate the adoption of the ECIT, we have developed a user-friendly version of the tool that can be accessed at: https://nationalethicsproject.org/course-id-tool. This version allows users to easily upload their catalogs in either comma delimited CSV or Excel format for analysis, and to receive as output a scored version of their catalog, with each course being scored with both the expert-driven and the data-driven ECIT. The scores are accompanied by a label indicating verbally the likelihood of a particular course being ethics-related (e.g., *"Almost certainly related to ethics", "May or may not be related to ethics")*.

Conclusion

For researchers and practitioners, the ECIT provides an efficient method for systematically identifying ethics-relevant courses, bringing into greater clarity where and how ethics is being addressed across a course catalog even when explicit use of the term "ethic" or "moral" is lacking. This method operationalizes a more nuanced and comprehensive definition of ethics education than existing methods that depend on fewer search terms. In addition, the use of the ECIT is supported by direct evidence that it yields results consistent with those produced by expert human judges, who demonstrated considerable agreement regarding the ethical relevance of a range of courses. For researchers, the tool can support more inclusive studies of pedagogy and learning. For administrators and faculty, it can be used to identify assets and to target resource allocation for building a coherent ethics education strategy that is clearly aligned with an institution's mission.

Acknowledgements

We are grateful to Ka Ya Lee and Drew Chambers for their careful review of the manuscript and for their valuable insight. We also extend our thanks to Cara Biasucci and Deni Elliott for assistance in recruiting participants. Their contributions were invaluable to the success of this project. This work was generously supported by the Edmond J. Safra Philanthropic Foundation and Eugene P. Beard.

References

American Association of Colleges and Universities. (2022, August 8). Essential learning outcomes. AAC&U. Retrieved September 19, 2022, from https://www.aacu.org/trending-topics/essential-learning-outcomes

Araque, O., Gatti, L., & Kalimeri, K. (2020). MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based Systems*, *191*, 105184.

Bankhead, M., Choi, Y. J., Patil, Y., & Thoma, S. J. (2022). Individual and School Correlates of DIT-2 Scores Using a Multilevel Modeling and Data Mining Analysis. *Applied Sciences*, *12*(9), 4573.

Barnwell, D. (1989). 'Naïve'native speakers and judgements of oral proficiency in Spanish. *Language Testing*, *6*(2), 152-163.

Bebeau, M. J., & Monson, V. E. (2008). Guided by theory, grounded in evidence: A way forward for professional ethics education. In *Handbook of Moral and Character Education* (pp. 573-598). Routledge.

Beever, J., Kuebler, S. M., & Collins, J. (2021). Where ethics is taught: An institutional epidemiology. *International Journal of Ethics Education*, *6*(2), 215-238.

Berry, J. W. (1989). Imposed etics—emics—derived etics: The operationalization of a compelling idea. *International Journal of Psychology*, *24*(6), 721-735.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21-41

Christensen, L. J., Peirce, E., Hartman, L. P., Hoffman, W. M., & Carrier, J. (2007). Ethics, CSR, and sustainability education in the Financial Times top 50 global business schools: Baseline data and future research directions. *Journal of Business Ethics, 73*(4), 347-368.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Colby, A., Ehrlich, T., Beaumont E., & Stephens, J. (2003). *Educating citizens: Preparing America's undergraduates for lives of moral responsibility*. JosseyBass.

Drechsler Sharp, M., Komives, S. R., & Fincher, J. (2011). Learning outcomes in academic disciplines: Identifying common ground. *Journal of Student Affairs Research and Practice, 48*(4), 481-504.

Elliott, D. (2007). *Ethics in the First Person: A Guide to Teaching and Learning Practical Ethics*. Rowman & Littlefield.

Elliott, D., & June, K. (2018). The Evolution of Ethics Education 1980–2015. In *Ethics Across the Curriculum—Pedagogical Perspectives* (pp. 11-37). Springer, Cham.

Finelli, C.J., Holsapple, M.A., Ra, E., Bielby, R.M., Burt, B.A., Carpenter, D.D., Harding, T.S., & Sutkus, J.A. (2012). An assessment of engineering students' curricular and co-curricular experiences and their ethical development. *Journal of Engineering Education, 101*(3), 469-494.

Fischman, W., & Gardner, H. (2022). *The real world of college: What higher education is and what it can be*. MIT Press.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, *50*(1), 344-361.

Glanzer, P. L., & Ream, T. C. (2008). Educating different types of citizens: Identity, tradition, and moral education. *Journal of College and Character, 9*(4), 1-14.

Griffith, S.M., Domenech Rodriguez, M.M., & Anderson, A.J. (2014). Graduate ethics education: A content analysis of syllabi. *Training and Education in Professional Psychology, 8*(4), 248-252.

Hamer, R. M. (1991). *A SAS macro for computing intraclass correlation coefficients*. Richmond: Virginia Commonwealth University.

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2018). Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, *4*(1).

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., ... & Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057-1071.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537-553.

Kidd, D., Miner, J., Schein, M., Blauw, M., & Allen, D. (2020). Ethics across the curriculum: Detecting and describing emergent trends in ethics education. *Studies in Educational Evaluation*, *67*, 100914.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Lambeth, E. B., Christians, C. G., Fleming, K., & Lee, S. T. (2004). Media ethics teaching in century 21: Progress, problems, and challenges. *Journalism & Mass Communication Educator, 59*(3), 239-358.

Matchett, N. J. (2008). Ethics across the curriculum. *New Directions for Higher Education*, *142*, 25-38.

May, D. R., Luth, M. T., & Schwoerer, C. E. (2014). The influence of business ethics education on moral efficacy, moral meaningfulness, and moral courage: A quasi-experimental study. *Journal of Business Ethics, 124*(1), 67-80.

Mayhew, M. J., Rockenbach, A. N., Bowman, N. A., Seifert, T. A., & Wolniak, G. C. (2016). *How college affects students: 21st century evidence that higher education works* (Vol. 3). John Wiley & Sons.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Yoshua Bengio et al., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mills, A., Sanders, A. K., & Hussain, S. S. (2019). Fitting it all in? A census of undergraduate ethics and leadership courses in accredited US journalism and mass communication programs. *Journalism & Mass Communication Educator, 74*(3), 265-275.

Neill, M. S. (2017). Ethics education in public relations: Differences between stand-alone ethics courses and an integrated approach. *Journal of Media Ethics, 32*(2), 118-131.

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *Observer, 31*(3), 19-21.

Pike, K. L. (2015). *Language in relation to a unified theory of the structure of human behavior* (Vol. 24). Walter de Gruyter GmbH & co KG.

Schouten, G. (2022). The case for egalitarian consciousness raising in higher education. *Philosophical Studies*, 1-24.

Triandis, H. C., Malpass, R. S., & Davidson, A. R. (1973). Psychology and culture. *Annual Review of Psychology, 24*(1), 355-378.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12.

Wang, L. C., & Calvano, L. (2015). Is business ethics education effective? An analysis of gender, personal ethical perspectives, and moral judgment. *Journal of Business Ethics, 126*(4), 591-602.

Weston, A. (2006). *A Practical Companion to Ethics, 3rd edition.* Oxford University Press.

You, D., Ruiz, A., & Warchal, J. (2018). The presentation of ethics in undergraduate psychology syllabi: A review. *Teaching of Psychology, 45*(4), 346-350.

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine, 31*(29), 3972-3981.

Table 1

Frequence of Course Catalogs Representing Carnegie Classifications

Carnegie Classification [classification #]	Population	Sample
Doctoral Universities: Very High Research Activity [15]	131	3
Doctoral Universities: High Research Activity [16]	135	2
Doctoral/Professional Universities [17]	152	2
Master's Colleges & Universities: Larger Programs [18]	350	3
Master's Colleges & Universities: Medium Programs [19]	196	3
Master's Colleges & Universities: Small Programs [20]	139	1
Bacc.Colleges: Arts & Sciences Focus [21]	241	3
Bacc. Colleges: Diverse Fields [22]	334	4
Bacc./Assoc.'s Colleges: Mixed Bacc./Assoc.'s [23]	151	1
Bacc./Assoc.'s Colleges: Assoc.'s Dominant [14]	111	1

Study	*Block*	*Raters*	*ICC* (Single rater)	*ICC* (*k* raters)
Study 1 *n* = 61	1	15	.67	.96
	2	17	.74	.98
	3	11	.40	.88
	4	10	.57	.93
	5	4	.59	.85
	6	4	.53	.82
Study 2 *n* = 12	1	3	.60	.82
	2	3	.70	.87
	3	3	.71	.88
	4	3	.65	.85

Table 2. ICC = interclass correlation coefficient.

Study	*Course group*	*M*	*SD*	*Minimum*	*Maximum*
Study 1	High-high	3.81	0.89	1.66	5
	Low-low	1.67	0.80	1	3.33
	High-Low	2.50	1.04	1	4.33
	Low-High	3.16	0.82	1	5
Study 2	High-High	3.83	0.60	2.66	5
	Low-Low	1.42	0.71	1	4
	High-Low	2.43	0.61	1	3.33
	Low-High	3.04	0.86	1	4.66

Table 3. High-high: course with high scores on both the data-driven and the expert-driven ECIT. Low-low: course with low scores on both the data-driven and the expert-driven ECIT. High-low: course with a high score on the data-driven ECIT and a low score on the expert-driven ECIT. Low-High: course with a low score on the data-driven ECIT and a high score on the expert-driven ECIT.

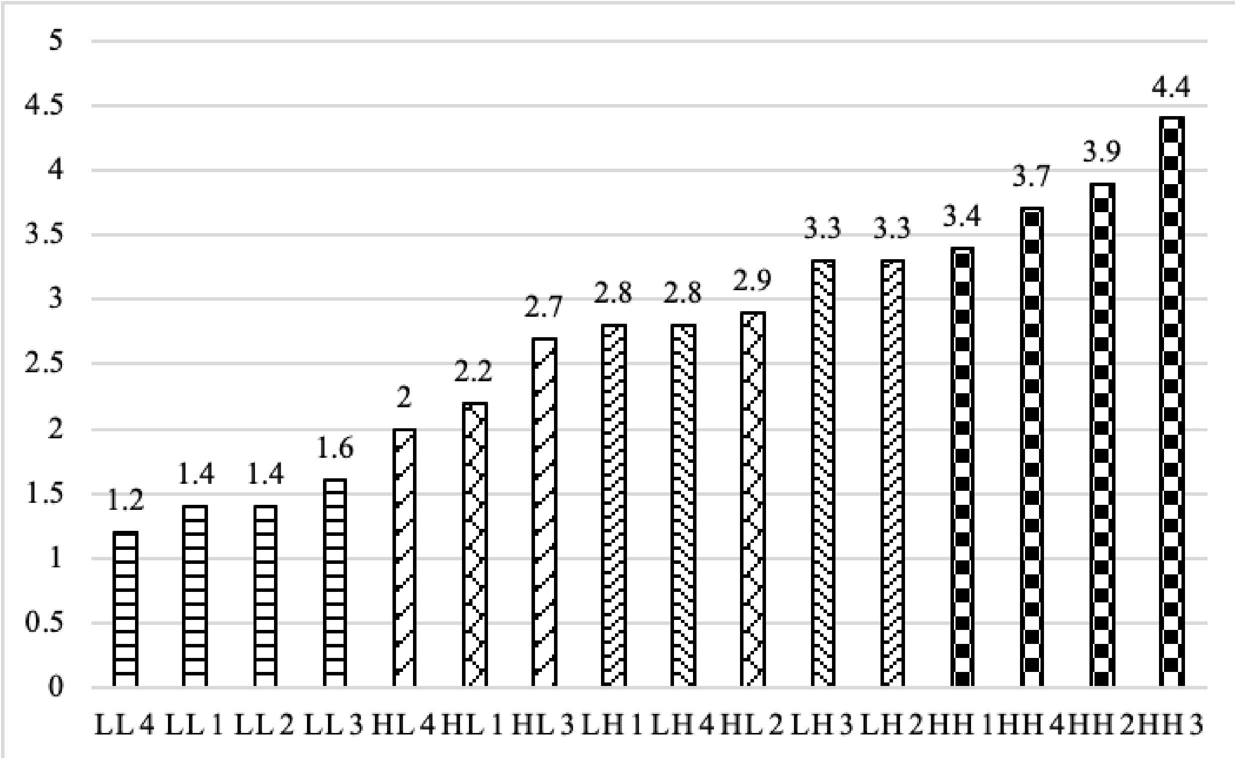

Figure 1. Ratings of course relatedness to ethics by ECIT category and block in Study 2. HH: course with high scores on both the data-driven and the expert-driven ECIT. LL: course with low scores on both the data-driven and the expert-driven ECIT. HL: course with a high score on the data-driven ECIT and a low score on the expert-driven ECIT. LH: course with a low score on the data-driven ECIT and a high score on the expert-driven ECIT.