Centre of Excellence for Development Impact and Learning

March 2022

Design Paper 11

Enhancing evaluation use: An experimental assessment of rubric methodology on collective judgements in Myanmar

Christina Peterson, Sten te Vogt and Jasmine Lal Fak Mawii

LIDC London International Development Centre

UKaid from the British people

Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL design paper

CEDIL design papers present the protocols of studies supported by the CEDIL programme. They describe the objectives, the methodology and the implementation plan of impact evaluations, evidence syntheses, and methodological papers supported by CEDIL.

About this design paper

This design paper was submitted to CEDIL by the "Enhancing Evaluation Use: Rubric Methodology in Program Evaluation" Project S.176 team.

Suggested citation: Peterson, C., te Vogt, S. & Lal Fak Mawii, J. (2022) 'Enhancing evaluation use: An experimental assessment of rubric methodology on collective judgements in Myanmar', CEDIL Design Paper 11. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford.

This project was funded by the Centre of Excellence for Development Impact and Learning (CEDIL), supported by UK aid from the UK Government. The views expressed in this research project paper do not necessarily reflect the UK Government's official policies or CEDIL.

Enhancing evaluation use: An experimental assessment of rubric methodology on collective judgements in Myanmar

Christina Peterson[1], Sten te Vogt[2], and Jasmine Lal Fak Mawii[3]

[1] University of Tennessee Knoxville, Department of Educational Psychology and Counseling
[2] Mekong Economics Myanmar
[3] Mekong Economics Myanmar

Abstract

Rubric methodology in program evaluation provides explicit criteria for synthesizing evidence and guiding collective judgements. Despite expectations and generally positive reflections, systematic research on the motivational, attitudinal, and relational influences program evaluation rubrics use has on collective judgements is scarce. Although rubric use in evaluation practice is growing worldwide, it is also unclear how this method for making evaluative judgements transfers across cultures since it was developed within a Western epistemological context. In this article, we outline the research protocol for an experiment aiming to assess how rubrics influence information sharing, preference effects, and ultimately collective judgements among groups within the cultural context of Myanmar. Our findings will highlight how rubric methodology transfers across evaluation contexts and provide insight on the conditions in which rubric methodology may be a useful approach for evaluation practitioners. Finally, we discuss the policy relevance of this research and alignment with CEDIL programme goals.

Keywords

Rubric, collective judgement, evaluation use, evaluation capacity building, culturally responsive evaluation

Evaluation capacity building (ECB) is a key component of evaluation strategies that aims to facilitate stakeholder participation in evaluation processes to improve stakeholder capacity for evaluative thinking and ability to make evaluative judgements about the merit, worth, and significance of a program (Buckley, Archibald, Hargraves, & Trochim, 2015). Strasser and Dietz-Uhler (2001) use the term "collective judgement" to refer to evaluative tasks undertaken among groups. Collective judgements differ from individual in the sense that groups comprise between individual preferences or use a different process of decision making than they would have as individuals. One tool, or "rule", that has been used to aid evaluative thinking and collective judgements among stakeholder groups is program evaluation rubrics. Yet, despite expectations and generally positive reflections (King & Allan, 2018; King & Guimaraes, 2016), systematic research on the motivational, attitudinal, and relational influences program evaluation rubrics use has on collective judgments is scarce (Peterson & Skolits, 2020; Tai, et. al., 2018). Preference effects, or an individual's propensity to remain committed to an initial preference or judgement despite evidence supporting alternative options, and conditions that foster information sharing are two such aspects that deserve additional examination in the context of evaluation rubric use. We propose that the effect of rubrics on information sharing and initial preferences mediates how the use of rubrics influences collective judgements. Although rubric use in evaluation practice is growing worldwide (King & Allan, 2018; King & Guimaraes, 2016; Tremblay, Bertrand, & Fraser, 2017), it is also unclear how this method for making evaluative judgements transfers across cultures since it was developed within a Western epistemological context. Therefore, this study aims to understand how rubrics influence social information sharing, preference effects, and ultimately collective judgements among groups within the cultural context of Myanmar.

Rubric Methodology

An evaluation rubric is a set of instructions or rules that aid evaluative judgments regarding the quality of an object. Rubrics are composed of three components: criteria, standards or scoring strategy, and descriptors (Martens, 2018). Criteria are the standard by which something may be judged or decided. In a program evaluation, these criteria might consist of program relevance, effectiveness, sustainability, or equity. The performance standards are a continuum of labels for levels or ranks from poor to excellent, while descriptors provide examples of what the program

must demonstrate in order to attain a particular performance level for each criterion. In the program evaluation context, rubrics are an analytical rule to help synthesize evidence and values for the purpose of making judgements regarding the merit and worth of program outcomes (Davidson, 2005). By rule, we mean a device for categorization, estimation, paired comparisons, or other judgmental tasks that require inference beyond the information given. In this way, rubrics are a "rule" that enable quantitative data to be interpreted alongside qualitative data, which can encourage methodological responsiveness to the program context and stakeholder values (Davidson, 2005; King, 2016). Rubrics help address the persistent challenge to evaluative thinking of synthesizing substantial amounts of information (see Scriven, 1995) because they establish quality criteria that guide evidence synthesis across multiple data sources (Davidson, 2005). By providing explicit quality criteria, Davidson (2005) argues evaluation rubrics may also help improve transparency and credibility of evaluative judgements.

Evaluative Judgements

In an ECB or participatory evaluation context, evaluative judgements are negotiated among a group of stakeholders who differ with respect to prior information, pre-existing preferences about the program, power positions, and strategic motivations. An evaluative judgement is conceptualized as the act of processing a stimulus and categorising it as "good" or "bad". This processing can occur intuitively (unconsciously), where evaluative judgements are often activated without one being aware they are categorising an object, or deliberatively (consciously) (Bargh & Chartrand, 1999). Deliberate processing is often referred to as "evaluative thinking" where individuals or groups are "motivated by an attitude of inquisitiveness and a belief in the value of evidence, that involves identifying assumptions, posing thoughtful questions, pursuing deeper understanding through reflection and perspective taking, and informing decisions in preparation for action" (Buckley, Archibald, Hargraves, & Trochim, 2015, p.378). Tasks that call for evaluative judgements differ from intellectual tasks (e.g., solving a math problem) in that they do not have a demonstrably correct answer (Laughlin & Ellis, 1986). Individuals bring pre-existing knowledge and preferences into the judgment task where evidence is either commonly known among stakeholders or uniquely known to make a collective judgement. Evaluation rubrics are meant to motivate evaluative thinking and deliberate judgements.

Evidence Sharing

The amount of information collected in an evaluation is often substantial, demanding considerable cognitive energy to understand, synthesize, and interpret. Lu, et al., (2012) observed that as the total amount of information available increases, so does the tendency to focus on information that is commonly known. Evidence suggests that groups disproportionately discuss information that is commonly known among group members more often than information that is known by only some group members to the detriment of subsequent decisions (Lu, Yuan, & McLeod, 2012). When judgments are made among a group, the task may also be characterized along a continuum from purely cooperative to purely competitive. If stakeholders perceive determining a judgment as a competitive task, there is a greater tendency to strategically withhold uniquely known information (Wittenbaum, Hollingshead & Botero, 2004). Laboratory studies find that, on average, about half of the shared and two-thirds of the uniquely held information goes unmentioned during discussions (Oliver, Hollingworth & Briner, 2015). Groups tend to share more uniquely held information under conditions where the task required selecting a logically preferred outcome compared with choosing between equally attractive solutions. Additionally, tasks in which there are fewer alternatives to choose between (discussing only two alternatives rather than three or more) or discussion time was less than 30 minutes tend to promote sharing uniquely held information (Oliver, Hollingworth & Briner, 2015). Yet, even when a facilitator attempts to draw out unique information, groups still generally discuss more commonly held information and arrive at sub-optimal collective judgements (Lu, Yuan, & McLeod, 2012).

Preference Effects

Evaluative thinking implies that stakeholders are able to suspend automatic judgements or be open to re-evaluating initial judgements. Once an initial judgment is activated, people tend to bias interpretation of subsequent information to support this initial judgement (Greitemeyer & Schulz-Hardt, 2003). Such information biasing leads to preference effects where sub-optimal preferences are maintained despite evidence to the contrary (Greitemeyer & Schulz-Hardt, 2003). The distribution of initial preferences individuals bring to a group discussion generally predicts collective judgements since prior knowledge and opinions have a strong normative influence on groups (Stasser & Dietz-Uhler, 2001). Sharing new information and synthesising it once mentioned during discussion is critical for releasing individuals from their initial,

presumably less informed, judgement (Greitemeyer & Schulz-Hardt, 2003; Stasser & Birchmeier, 2003). Although intuitive, or automatic, judgements based on heuristic rules perform as well or better than judgements based on conscious evaluative thought in many contexts (Gigerenzer & Brighton, 2009), they may not be accepted as credible forms of judgement in the evaluation context. Among stakeholders charged with making decisions about social programs, the transparency and credibility of collective judgements are important criteria for evaluation use. We propose that rubrics are a method for synthesizing information that increases information sharing and reduces preference effects to improve credibility and transparency of collective judgements (Figure 1).

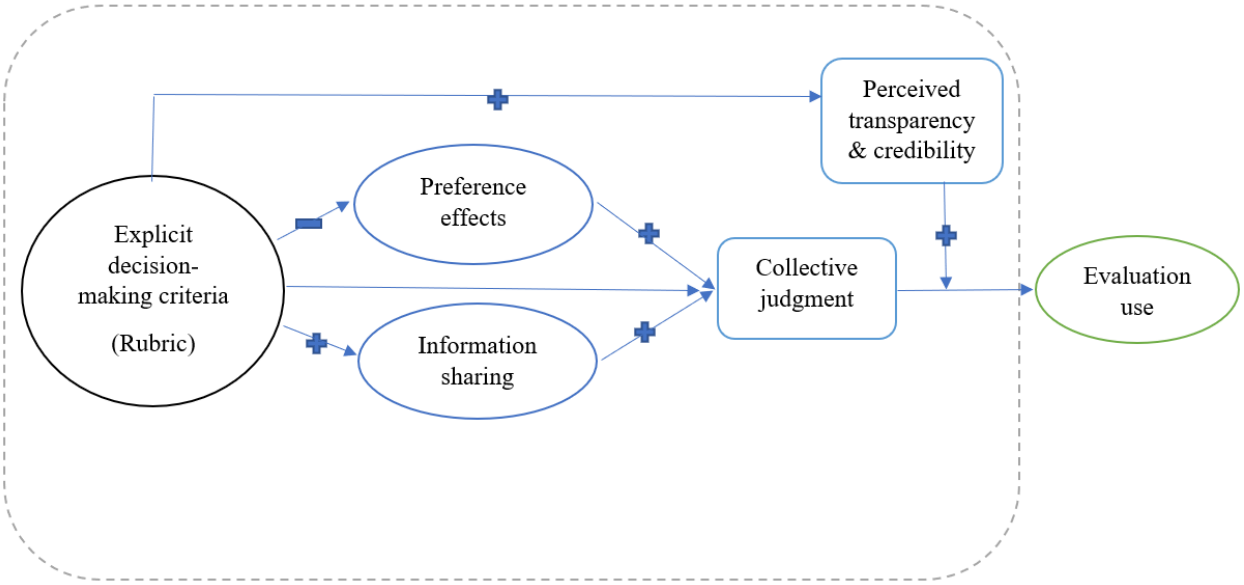

Figure 1: Conceptual model of rubric influence on evaluative judgments and evaluation use

Myanmar Context

Evaluation capacity building (ECB) among local support NGOs is a priority for monitoring and evaluation support activities in Myanmar (Van Hemelrijck, 2017). The transferability of ECB activities, particularly rubrics, to the Myanmar context requires attention to cultural values and epistemologies. Several studies have examined the cultural characteristics generally exhibited in Myanmar and suggest a context characterized by moderately high tendencies toward collectivism and uncertainty avoidance with a moderate long-term orientation (Minkov, et. al, 2017; Rarick &

Nickerson, 2006). High power distance, or a propensity to value unequal or hierarchical social relations, has been observed in several studies and considered a barrier to approaches, such as student-centered instructional methods (Tyrosvoutis, 2016; Rudkin & Erba, 2018). Moreover, Myanmar participants also exhibit greater intrapersonal and interpersonal emotional competence and favor conflict styles that place a high concern for the other party (Min & Takai, 2018), which may influence information sharing or perceived decision transparency and credibility among groups. These cultural norms may mean that Myanmar stakeholders accept and expect unequal power distributions in an evaluation setting and pay more attention to relational factors in order to maintain interpersonal harmony when making collective judgements.

Epistemological orientations influence approaches to scientific and philosophical questions (Nisbett, Peng, Choi, & Norenzayan, 2001).Western notions of ECB implicitly value debate of evidence among participants of differing social positions and roles. Yet, in cultural contexts, like Myanmar, where in-group harmony may be prioritized over individual agency, any form of confrontation, such as debate, may feel uncomfortable or even be discouraged (Nisbett, Peng, Choi, & Norenzayan, 2001). Variations in reasoning styles stemming from these epistemological foundations have also been observed (Nisbett, Peng, Choi, & Norenzayan, 2001). Attempts to categorize objects (i.e. good program or poor program) may not be an important epistemic goal (Nisbett, Peng, Choi, & Norenzayan, 2001). As such, many Asian cultures are more inclined think more holistically and attribute causation to contextual conditions. In contrast, Westerners are often more analytical and reductionist, attributing causality to isolated factors (e.g. the program) (Nisbett, Peng, Choi, & Norenzayan, 2001; Sanchez-Burks, et al., 2003).

Research from a Western perspective suggests that people make better deliberate (conscious) decisions when there are explicit rules or criteria (Dijksterhuis & Nordgren, 2006). This is consistent with analytical reasoning styles (i.e. linear, discrete, less tolerant of contradiction) observed in Western cultures, making program evaluation rubrics appealing because they create explicit rules and weighting criteria, making conscious thought (evaluative thinking) easier. Although rubrics may offer a way to merge reasoning styles in a way acceptable to stakeholders of diverse cultural backgrounds by positioning both contextual information (e.g. qualitative or experience-based) alongside decontextualized, rule-based knowledge (e.g. surveys,

quasi-experimental methods), the need to examine the transferability of evaluation rubrics to a non-Western context like Myanmar is evident.

Methods

This study uses an experimental approach to investigate how program evaluation rubrics influence evaluative judgements and decision transparency and credibility within a hidden profile paradigm (Toma & Butera, 2009; Van Swol, Savadori, & Sniezek, 2003). The hidden profile paradigm of research refers to tasks in which individuals within a group are each privy to a unique set of commonly known and uniquely known information about a target object. The correct solution cannot be detected by the individual prior to group discussion. Results have consistently shown that groups focus their discussion on commonly known information and fail to arrive at optimal judgements when uniquely held information remains unshared (Lu, Yuan, & McLeod, 2012; Stasser & Titus, 1985). An experimental approach is appropriate in this study to reduce threats to internal validity and isolate the causal effect of rubrics on theoretical outcomes. To separate the effect of rubrics as an analytical rule from the participatory process typically used to generate them in program evaluation, this investigation uses rubrics that were not created by study participants. The following hypotheses will be tested:

> H1: Hidden profile groups using evaluation rubrics make significantly different judgments than hidden profile groups without rubrics.
>
> H2: The effect of rubrics on group judgements is mediated by preference effects and amount of information sharing.
>
> H3: Hidden profile groups using evaluation rubrics perceive group judgments as more credible and transparent than hidden profile groups without rubrics.

Design. The study will use a randomised block design. Participants will be recruited from several classes at a university in Yangon, Myanmar. To control for variation induced by student self-selection into different courses, class will be treated as a random blocking factor to increase analysis accuracy. Each participant will be randomised into one of two treatment conditions:

hidden information and rubric; hidden information without a rubric, Groups without a rubric will receive the 'next best alternative' treatment. In this case, the next best alternative is evaluative program performance based its intended goals alone.

Task. This experiment uses a rating task to test the aforementioned hypotheses. Previous research in the hidden profile paradigm typically uses selection tasks (see Greitemeyer & Schulz-Hardt, 2003; Greitemeyer, SchulzHardt, Brodbeck, & Frey, 2006). Rating tasks require participants to rate (e.g., "good" or "poor") a single object whereas selection tasks ask participants to select one (e.g., "the best") of several objects. Hidden profiles will be created about a water, sanitation, and hygiene program. Profiles will be constructed so that evidence supportive of judgement (C) is "hidden" (i.e. not all group members have this information included in their profile), and the judgements appearing to be best supported before discussion (A, B, or D) are most salient to individuals. Individuals will each be asked to rate the program as excellent, good, adequate, or poor before the group discussion. Then, groups will be asked to collectively rate the program using a rubric and program goals or program goals alone (i.e. next best alternative).

***Subjects*.** Myanmar participants studying for a master's degree in international development, public administration and similar fields from classrooms within a university in Yangon will be recruited to participate in the study. We conducted a power analysis based upon a chi-square (df=6) using a large group level effect size (odds ratio=6.64) as observed in previous hidden profiles research (all conducted in the United States). Based upon these results, our estimated sample size to achieve a power of .80 is 108 individuals divided into 36 groups (18 groups per condition). This sample size is consistent with sample sizes used in previous studies in the hidden profile paradigm (Greitemeyer & Schulz-Hardt, 2003; Rijnbout & McKimmie, 2014).

Procedures. Students will be invited to participate in a "group activity to help evaluate a WASH program in Myanmar". Upon consenting to participate, they will be given a unique ID number and randomly assigned to a 3-member group. Then, groups will each be randomly assigned a treatment condition. Each person within the group will receive a one-page evaluation summary containing information about a social program (i.e., either a hidden profile or complete profile)

and their participant ID number (see Appendix A). The entire set of information will contain approximately 24 items: 15 shared and 9 critical unshared items, both qualitative and quantitative. A critical item is defined as an item that is necessary to forming a complete judgement. Critical items have a strong positive or negative valence rather than describing neutral background or contextual information. Positive, negative, and neutral information will be purposefully mixed together to form three distinct program profiles in which full information is "hidden" from each profile. To avoid order effects, the order of information presentation will be reversed for half the participants. Participants will first read their profile individually and take notes regarding the information that they may use in the group discussion. After 15 minutes, participants will be asked to individually make an initial judgement (i.e. poor, adequate, good, excellent) about the program and write it on their executive summary profile. Executive summary profiles will be collected by the research team.

After making an individual judgement, participants join their group for discussion. All treatment conditions will receive a form with a list of program objectives and a group ID number. In addition, the rubric treatment groups will receive a rubric. Groups assigned to different treatments will be isolated, so they do not know the conditions of other treatment conditions. Groups will be instructed to discuss the program and make a group judgment about the value of this program (i.e. poor, adequate, good, excellent). Participants will then report individually whether they agree with the group decision in a follow up survey. This survey will also assess perceived information credibility and decision-making transparency. Participant demographic information will be collected along with previous evaluation experience. Group sessions will be audio-recorded to assess the amount of uniquely held information that is shared. Each recording will be transcribed, coded and assessed.

Instrument development. The executive summary and rubrics are being developed de novo for use in both English and Burmese. Scales for measuring these credibility and transparency exist in English, so we will follow Sperber, Devellis and Boehlecke's (1994) method for cross-cultural instrument development using back-translation. For all instruments, the initial translation will be completed by a bilingual research assistant experienced in the international development field. The back-translation will be completed by a second bilingual research assistant who was not involved in the initial instrument construction. Research team members not involved in the

translation process will assess the original English versions with the back-translated versions on the criteria of language comparability and similarly in interpretation. Resulting instruments will be similar, yet culturally relevant, and pilot tested (n=12) prior to experiment implementation.

The first phase of instrument construction will use exploratory focus groups (FGDs) (k=2, n=12) to study the significance and meaning of information credibility and decision transparency among Burmese speaking participants recruited from local non-governmental organizations. In addition, the FGDs will examine the relevance of item content from established English language scales (de Fine Licht, 2011; Meyer, 1988). After analysing FGDs data, a team of Myanmar and Western researchers will construct instruments in Burmese based upon the English versions. All instruments (executive summary, rubrics, and credibility/transparency scale) will be pilot tested among bi-lingual participants in Myanmar. These participants (n=20) will be recruited from local non-governmental organizations and invited to the Mekong Economics office to participate in the pilot test. They will be asked to read the executive summary and rubrics, make a judgement about what they read (i.e. poor, adequate, good, excellent), and take the credibility/transparency survey. Half of the participants will receive English versions and half will receive Burmese versions. After completing these tasks, they will be asked for feedback and comments on the instruments and we will assess reliability between the two versions.

Analysis. The project uses a hypothesis-wise significance level of $\alpha = .05$. Missing data will be handled using Expectation Maximization methods if missingness is greater than 5% and missing at random (MAR). Using classroom as a blocking factor should increase statistical power and efficiency of our sample size. Prior to hypothesis testing, we will conduct a manipulation check. Using Chi-square analysis, we will validate that the distributions of pre-discussion preferences significantly differed across profile conditions (hidden_A, hidden_B, and hidden_C) to make sure experimental manipulation of preferences was successful. In the complete profile condition, information exchange should not alter the participants' preferences and we will use a test for similarity to assess the manipulation.

Hypothesis 1. The total effect of rubrics on group decision will be analyzed using a two-level generalized hierarchical model with a probit link function. Discussion time (minutes) will be included as a covariate and classroom as a level two random factor.

Hypothesis 2. The relationship between rubrics, information sharing, preference effects, and group decision will be analyzed using a three-level generalized hierarchal regression (probit link function) with parallel mediation effects for preference effect and information sharing. Preference effect is a level 1 variable coded as either 0 (no change) or 1 (change) for each participant. Information sharing is a level two group variable proportional outcome bounded between 0 and 1. Discussion time (minutes) and treatment condition will also be included as a level two covariates. Classroom will be included in the model as a level three random factor. To generate accurate standard errors and compare indirect effects in a multiple mediator model, we will use a bootstrapping technique (Preacher & Hayes, 2008).

Hypothesis 3. The effect of rubrics on information credibility and transparency will be analyzed using a three-level hierarchical linear model, where credibility and transparency will be individual scores taken from scales with a 5-7-point response structure. Group decision, discussion time (pending significance in analysis one), and treatment condition will be included as level two covariates. Classroom will be included in the model as a level three random factor.

Table 1: List of variables included in analysis

Level	Outcome	Dependent Variable	Variable Type
Individual	Preference effect	Pre-discussion judgement; Post-discussion judgement	Binary (change, no change)
	Information credibility	Post-discussion perceived credibility	Likert-Continuous
	Decision transparency	Post-discussion perceived transparency	
Group	Collective judgment	Group judgment	Multinomial
	Information shared	Proportion of common and unique information shared in discussion	Proportion

Time to decision	Minutes to group decision	Continuous

Limitations

This experiment has several notable threats to construct validity, external validity, internal validity, and statistical conclusion validity. Construct validity deals with the relationships between construct operationalization and measurement process (Chen & Rossi, 1987). In operationalizing information credibility or decision transparency, measurement error may occur due to response imprecision or cultural irrelevance. Even for more direct measurement of information sharing and preference effects, there is potential for measurement error or inexact overlap between the theoretical constructs and the measurements used. In regard to rubrics, the level of detail in criteria descriptions matters, with more detail leading to more extreme judgments (Van Boven & Epley, 2003). As such, how the rubric is constructed will influence measurement of groups judgment in this study.

External validity refers to the "approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times" (Cook & Campbell, 1979, p. 37). The ability to generalize results from this study to other contexts is limited because groups ultimately bare no accountability for the judgments they make. Not being accountable for the judgment may induce different behavior than what would be observed in an evaluation setting. Participants are not affiliated with the program so their initial judgements may be more amenable to change than if they had more of a stake in the program. They are also less constrained by political factors that influence judgements and propensity to share information. Despite these limitations, we mitigate threats to external validity by recruiting participants who have experience in international development. Conceivably, most of these participants are familiar with the logic of program evaluation.

This study may also be impacted by Hawthorne effects in which participants act differently than they normally would because they know they are under investigation. However, the Hawthorne effect may actually mimic the evaluation context where stakeholders may feel "observed" by external evaluators. Finally, this study captures a one-shot interaction between group members, which may not reflect all evaluation contexts. The experiment takes place on a

university campus, a setting with which participants are presumably familiar and comfortable, reducing the likelihood of unfamiliar environmental cues influencing participant behavior. Moreover, it is not implausible that a stakeholder meeting to discuss evaluation findings would occur in a conference room at a university.

Cook and Campbell (1979, p. 37), internal validity "refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause". In other words, was it really the rubrics that influenced the (significant) result of this experiment? Cook and Campbell (1979) outline four possible threats to internal validity that cannot be controlled with randomization. The first is differential attrition from experimental groups due to differences in demands made on subjects within a group. Groups receiving a rubric may find that it is difficult or cognitively demanding to use, resulting in individuals deciding to leave the experiment early. Second, systematic bias could occur in the rubric groups if individuals are confused with the task or use the rubrics in unanticipated ways. If significant confusion occurs, results could be the effect of an unidentified heuristic groups used to compensate for not understanding how to use the rubric, rather than rubric.

Statistical conclusion validity refers to the adequacy of the statistical model and power to detect significant results that actually exist. Although this study uses randomization, it does not eliminate the effects of all external influences on the outcomes of interest. Controlling for such variables reduces error variance and confidence intervals but requires more complicated statistical models. Using hierarchical linear models has the advantage of incorporating mediators and the classroom blocking factor to improve model specification and power. Currently, the literature does not fully address the relationship between rubric use, discussion time, and collective judgment. Therefore, our use of discussion time as a covariate rather than a mediator or moderator could be an incorrect model specification, leading to biased coefficients or larger error variance. Although we did not power this experiment on a moderation analysis, additional exploratory analysis will be included to offer recommendations for future research.

Policy Relevance & Innovation

There is growing recognition in the evaluation field that basic research on evaluation needs to be prioritised so that evaluators have access to evidence-based practices to improve evaluation

impact. A survey of American Evaluation Association (AEA) members identified that evaluators believe research on the impact of evaluation, evaluation methods, and evaluation contexts are top priorities in the field (Szanyi, Azzam, & Galen, 2012). The proposed study addresses questions related to both research on methods (i.e. How do different stakeholders view the credibility of different methodological approaches?) and research on context (i.e. What contextual factors alter the evaluation? How do they affect evaluation methods?) to help evaluators select appropriate practices and advance evaluation use among local stakeholders.

Rubrics are also a promising strategy for building evaluative thinking capacity more broadly. Reflections by Tremblay, Bertrand, and Fraser (2017) suggest rubrics methods perform well in evaluation contexts where there is sufficient time to develop valid rubrics ahead of data collection, evaluators are experienced facilitators, and clients are willing to be involved. King, McKegg, Oakden, and Wehipeihana, (2013), add that contexts in which stakeholders are prepared to openly discuss and debate values and there are a diverse range of competing stakeholder priorities, perspectives and values are also necessary for effective rubric use. This study contributes to understanding how rubric use influences decisions in contexts where there is significant information asymmetry and pre-existing preferences about a program among stakeholders. Significant results would indicate rubrics are an appropriate method for making evaluative judgements under these conditions.

The proposed project also aligns with the priorities of CEDIL to develop and adapt evaluation methods and ensure DFID can provide guidance on strategies for addressing complex evaluation challenges. One of the challenges identified is the gap in the methods for jointly drawing upon evidence from process and impact evaluations to answer evaluation questions (Oliver, Gough, Copestake, & Thomas, 2018). Furthermore, the integration of qualitative and quantitative information from mixed-method designs poses a persistent problem in making evaluative judgements. Evaluation rubrics offer one strategy for addressing this gap because they provide a transparent analytical framework for synthesising numerous data sources. Such knowledge can guide evaluators in methods of synthesising evidence to ensure that inference is transparent and replicable.

APPENDIX A: Draft Executive Summary

(1) International donors, in collaboration with local development organizations, implemented a three-year water, sanitation and hygiene (WASH) program. (2) Donors invested 5 million (USD) on water infrastructure, operational costs, and data collection to reduce health risks by improving access to safe sustainable water supplies, sanitation facilities and promoting hygiene practices. (4) This evaluation aims to assess program relevance and reach, effectiveness, sustainability, and equity using surveys and focus groups.

Relevance & Reach

(3) The WASH program benefits more than 70,000 people in 48 villages, including 20 schools. (19) The WASH program activities were very relevant to people's health and livelihood needs. (20) Among villagers, 96% were able to discuss good practice and knowledge about WASH programs. (21) The program also invested in water systems that align with existing local government priorities and increases sustainability. (22) Many local government officials said they have improved their knowledge and capacity in community-based programming, technical WASH as well as program and financial management.

Effectiveness

(5) Most villages use the cleaner water from the project's systems for consumption and food preparation, while water from other sources is used for clothes washing, general cleaning, and bathing. (6) The people and authorities interviewed believe the project reduced seasonal outbreaks of diarrhea among children. (7) Survey data suggests incidence of diarrhea decreased by 67%. (8) However, the survey noted that only 30% of beneficiaries can access water within 500 meters. (9) In addition, only 24% of women report saving time and money getting dry season (drinking) water via newly installed pipes (10) Moreover, on average, women saved just 12 minutes per day collecting water.

Sustainability

(13) The remote areas of the project intervention made implementation more time-consuming and not cost-effective. (14) This resulted in a project in which more than half (52%) of the budget was spent on operational overhead rather than community needs. (15) The evaluation shows that the WASH program successfully provided training for water system operation and maintenance for local government officials. (16) However, the program did not invest enough resources to support maintenance for breakages, replacement or damage to sustain water infrastructure. (17) The water systems are costly to maintain long-term and will only last 10-15 years with the current budget. (18) There also appears to be a chronic problem with the water infrastructure that leads to partial to total blockages of water flow.

Equity

(11) The team did not adapt the program to meet the needs of different sub-groups, resulting in only 42% of persons with a physical disability reporting that new latrines were accessible. (12) Thus, some individuals or groups were not served by this program. (23) The village committees consulted during the program implementation phase consisted of (73%) men and only (27%) of

women, which did not meet expectations for gender inclusion. (24) Nevertheless, 79% of women reported satisfaction with the WASH program overall.

APPENDIX B: Draft Rubric/Matrix

	Relevance and Reach	Effectiveness	Sustainability	Equity
Excellent	Project aligns with local government priorities Program reaches at least 70,000 people in 50 villages	90% of villagers articulate good WASH knowledge and practices Decrease diarrhea among children by at least 50% from baseline At least 80% of beneficiaries can access water within 500 meters	Local government officials have technical knowledge and budget to maintain infrastructure beyond 20 years 80% or more of program budget is spent on village needs and operational costs are less than 20% of budget	60% or more of women report saving time and money on water collection 90% of persons with disabilities report latrines are accessible
Good	Project aligns with local government priorities Program reaches at least 60,000 people in 40 villages	80% of villagers articulate good WASH knowledge and practices Decrease diarrhea among children by at least 40% from baseline At least 70% of beneficiaries can access water within 500 meters	Local government officials have technical knowledge and budget to maintain infrastructure 15-20 years 80% or more of program budget is spent on village needs and operational costs are less than 20% of budget	50% or more of women report saving time and money on water collection At least 80% of persons with disabilities report latrines are accessible
Adequate	Project doesn't align with local government priorities but meets a critical need in the village	70% of villagers articulate good WASH knowledge and practices Decrease diarrhea among children	Local government officials have technical knowledge and budget to maintain infrastructure 10-15 years	40% or more of women report saving time and money on water collection

	Program reaches at least 50,000 people in 30 villages	by at least 30% from baseline At least 60% of beneficiaries can access water within 500 meters	60% or more of program budget is spent on village needs and operational costs are less than 40% of budget	At least 70% of persons with disabilities report latrines are accessible
Poor	Project doesn't align with local government priorities or meet a critical need in the village Program reaches less than 50,000 people or less than 40 villages	Less than 70% of villagers articulate good WASH knowledge and practices Decrease diarrhea among children by less than 30% from baseline Less than 60% of beneficiaries can access water within 500 meters	Local government officials have technical knowledge and budget to maintain infrastructure less than 10 years Less than 60% of program budget is spent on village needs and operational costs exceed 40% of budget	Less than 40% of women report saving time and money on water collection Less than 70% of persons with disabilities report latrines are accessible

References

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W. M. (2015). Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation*, *36*(3), 375-388.

Chen, H. T., & Rossi, P. H. (1987). The theory-driven approach to validity. *Evaluation and program planning*, *10*(1), 95-103.

Cook, T. D., & Campbell, D. T. (1979). The design and conduct of true experiments and quasi-experiments in field settings. In *Reproduced in part in Research in Organizations: Issues and Controversies*. Goodyear Publishing Company.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks: Sage Publications.

de Fine Licht, J. (2011). Do we really want to know? The potentially negative effect of transparency in decision making on perceived legitimacy. *Scandinavian Political Studies, 34(3), 183-201.*

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological science, 1(2), 95-109.*

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, *1*(1), 107-143.

Greitemeyer, T., & Schulz-Hardt, S. (2003). Preference-consistent evaluation of information in the hidden profile paradigm: Beyond group-level explanations for the dominance of shared information in group decisions. *Journal of personality and social psychology*, *84*(2), 322.

King, J., & Allan, S. (2018). Applying evaluative thinking to value for money: The Pakistan Sub-national Governance Programme. https://www.nzcer.org.nz/system/files/journals/evaluation-maters/downloads/EM2018_207.pdf

King, J. & Guimaraes, L. (2016). Evaluating value for money in international development: The Ligada female economic empowerment programme. eVALUation Matters, Third Quarter, 2016. Africa Development Bank. http://idev.afdb.org/sites/default/files/documents/files/Evaluating%20value%20for%20money%20in%20international%20development-.pdf

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of MultiDisciplinary Evaluation*, *9*(21), 11-20.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177-189.

Martens, K. S. (2018). Rubrics in program evaluation. *Evaluation Journal of Australasia*, *18*(1), 21-44.

Meyer, P. (1988). Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly*, *65*(3), 567-574.

Min, M. C., & Takai, J. (2018). The effect of emotional competence on relational quality: Comparing Japan and Myanmar. *Current Psychology*, 1-11.

Minkov, M., Dutt, P., Schachner, M., Morales, O., Sanchez, C., Jandosova, J., Khassenbekov, Y., & Mudd, B. (2017). A revision of Hofstede's individualism-collectivism dimension. *Cross Cultural & Strategic Management*.

Oliver, S., Hollingworth, K., & Briner, R. (2015). Effectiveness and efficiency of committee work: a rapid systematic review for NICE by its Research Support Unit. *London: National Institute of Health and Care Excellence*.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological review, 108*(2), 291.

Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American psychologist, 54*(9), 741

Peterson, C., & Skolits, G. (2020). Value for money: A utilization-focused approach to extending the foundation and contribution of economic evaluation. *Evaluation and Program Planning*, *80*, 101799.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879-891.

Rijnbout, J. S., & McKimmie, B. M. (2014). Deviance in organizational decision making: using unanimous decision rules to promote the positive effects and alleviate the negative effects of deviance. *Journal of Applied Social Psychology*, *44*(7), 455-463.

Rarick, C. A., & Nickerson, I. (2006). An exploratory study of Myanmar culture using Hofstede's value dimensions. *Available at SSRN 1114625*.

Rudkin, A., & Erba, J. (2018). Myanmar's cultural dimensions: Exploring the relationship

among the social identity, attitudes towards globalisation and preferences of Myanmar consumers in Yangon. *International Journal of Asia Pacific Studies, 14*(1), 191–226.

Sanchez-Burks, J., Lee, F., Choi, I., Nisbett, R., Zhao, S., & Koo, J. (2003). Conversing across cultures: East-West communication styles in work and nonwork contexts. *Journal of personality and social psychology*, *85*(2), 363.

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation: methodology and validation. *Journal of cross-cultural psychology*, *25*(4), 501-524.

Stasser, G., & Dietz-Uhler, B. (2001). Collective choice, judgment, and problem solving. In: Hogg, M.A., Tindale, R.S. (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 31-55). Malden, MA: Blackwell.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education*, 1-15

Toma, C., & Butera, F. (2009). Hidden profiles and concealed information: Strategic information sharing and use in group decision making. *Personality and Social Psychology Bulletin*, *35*(6), 793-806.

Tremblay, G. H., Bertrand, F., & Fraser, M. (2017). Using Rubrics for an Evaluation: A National Research Council Pilot. *Canadian Journal of Program Evaluation*, *32*(2).

Tyrosvoutis, G. (2016). Taking the sage off the stage identifying obstacles to student-centered instruction on the Thai-Myanmar border. *International Education Journal: Comparative Perspectives*, *15*(4), 112-132.

Van Boven, L., & Epley, N. (2003). The unpacking effect in evaluative judgments: When the

whole is less than the sum of its parts. *Journal of Experimental Social Psychology*, *39*(3), 263-269.

Van Hemelrijck, A. (2017). Governance in Myanmar: Evaluation of the 'Building equitable and resilient livelihoods in the Dry Zone'project. https://www.lift-fund.org/sites/lift-fund.org/files/publication/er-governance-myanmar-effectiveness-review-240117-en%20%282%29.pdf

Van Swol, L. M., Savadori, L., & Sniezek, J. A. (2003). Factors that may affect the difficulty of uncovering hidden profiles. *Group Processes & Intergroup Relations*, *6*(3), 285-304.

Wittenbaum, G. M., Hollingshead, A. B., & Botero, I. C. (2004). From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. *Communication Monographs*, *71*(3), 286-310.

www.CEDILprogramme.org