Schendel, R. and Tolmie, A. (forthcoming) 'Beyond Translation: Adapting a performance-task-based assessment of critical thinking ability for use in Rwanda.' *Assessment and Evaluation in Higher Education*. DOI: 10.1080/02602938.2016.1177484.

Abstract

Critical thinking is frequently proposed as one of the most important learning outcomes of a university education. However, to date, it has been difficult to ascertain whether university students in low-income contexts are improving in their critical thinking skills, because the limited studies in this domain have relied on instruments developed in Western contexts, despite the clear dangers of such an approach. Cultural bias in assessment can best be overcome by explicitly developing tests for use in specific contexts. However, resource constraints often prevent this possibility. An alternative strategy is to adapt an existing instrument for use in a particular context. Although adaptation is the norm for high-stakes cross-cultural assessments, it is often not attempted for single country research studies. This may be due to an assumption that adaptation is excessively technical or will add significantly to a study timeline. In this article, which relies on data from a recent study in Rwanda, we present a methodology for adapting a performance-task-based assessment of critical thinking. Our experience with this methodology suggests that small teams can adapt instruments in a relatively short time frame, and that the benefits of doing so far outweigh any cost.

Keywords: Critical thinking; assessment; cross-cultural validity

Issues of cultural variance have long featured in the assessment literature. It is well documented that cultural background can have a significant impact on test performance, due to both explicit and implicit cultural nuances within test questions (Rogoff, 2003). As a result, cultural differences within samples can lead to construct-irrelevant variance in test scores unless efforts are made to ensure that all participants are equally likely to comprehend test questions in the same manner. Content familiarity is particularly important when assessing a complex ability, such as problem solving or critical thinking, given the necessity for test takers to engage both cognitive and metacognitive skills when responding to questions (Kuhn 1999; Mayer and Wittrock 1996). When faced with a familiar assessment scenario, participants are better able to quickly understand the content of test questions (Serpell 2007), thereby reducing cognitive load and allowing for the use of more complex cognitive strategies (Lun, Fischer, and Ward 2010). Assessment of 'higher order' thinking skills is, therefore, particularly challenging in cross-cultural contexts.

Although these challenges are frequently acknowledged by those seeking to assess complex cognitive abilities across multiple cultural contexts (e.g. the multi-country Programme for International Student Assessment, or PISA), to date, they have not been much discussed in reference to single country studies. However, the same problems emerge within single country studies that rely on instruments imported from other cultural contexts. Although the comparability of scores between different participants may be less of a challenge in single

country studies, comparability remains an issue when one seeks to use an imported instrument to assess participant ability to demonstrate a particular cognitive skill. Despite these concerns, the higher education literature is replete with studies adopting such an approach, particularly in low-income country contexts. For research teams operating in less-resourced environments, the development of new assessment tools is often not a feasible option, so there is a tendency to adopt tools validated elsewhere. Although imported assessments are typically translated into the local language before implementation, efforts to adapt the questions for cultural relevance are rare. This may be because of a lack of familiarity with adaptation methods, or because there is a fear that the process of adaptation will affect the psychometric properties of the instrument or will add excessively to a study timeline. Regardless of the cause, the net effect is that much of the literature on higher education learning outcomes in low-income contexts is reliant on assessments which have not been appropriately adapted and validated, which substantially limits the potential for research to inform higher education policy and practice.

This paper aims to contribute to this important gap in the literature by presenting a cultural adaptation methodology that was used in a recent study of one such complex cognitive skill – critical thinking – in Rwanda. We first present a justification for using the Collegiate Learning Assessment (CLA) as the underlying model for the adapted instrument before describing the adaptation methodology in detail. The article concludes with an evaluation of the benefits and challenges of using the methodology, thereby inviting reflection on its potential for use in other contexts. By presenting our experiences with this methodology, we intend to demonstrate that it is possible for small research teams to adapt instruments in a relatively short time frame and to argue that the benefits of doing so far outweigh any cost.

Background and rationale

In the context of the global 'knowledge economy', critical thinking is frequently proposed as one of the most important learning outcomes of a university education (Davies and Barnett 2015). In low-income contexts, critical thinking is often promoted as a key priority for international development, as the ability to think critically about problems and use evidence when making decisions is seen as a prerequisite for the adaptation of technology to local contexts and the proposal of new solutions to intractable challenges (Ashcroft & Rayner 2011; UNESCO 2009). Outside of the economic sphere, critical thinking is also viewed as crucial for global sustainability (Rieckmann 2012) - an issue which has increased salience in light of the new Sustainable Development Goals (United Nations 2015) – political participation (Luescher-

Mamashela et al. 2015) and broader conceptualisations of human development (Boni and Gasper 2012).

However, despite the perceived importance of critical thinking, it is not always clear that students are developing these skills during their time at university. Although evidence to the contrary has recently emerged in a number of high-income contexts (Arum and Roksa 2011; Blaich and Wise 2010; Phan 2011), to date, it has been difficult to verify whether such trends are also reflected in less-resourced settings. There have only been a few attempts to empirically assess the ability of university students to demonstrate critical thinking skills in low-income contexts (Bataineh and Zghoul 2006; Lombard and Grosser 2004, 2008; Osman and Githua 2009; Saavedra and Saavedra 2011), and all of these have relied on assessments developed for the U.S. market. As a result, very little is known about critical thinking in less-resourced parts of the world, as the poor scores obtained within the existing study samples may simply indicate limited understanding of the cultural frame of reference underpinning the imported assessments.

In 2011, one of the authors set out to address this gap in the literature by investigating the extent to which Rwandan undergraduates are improving in their critical thinking ability during their time at university (Schendel 2015). As the study design required a large-scale quantitative assessment of student critical thinking ability, it was necessary to identify an assessment tool that would be valid for use in the study context. However, no assessment tool had ever been tested for use in Rwanda, and the resources available for the study prevented the possibility of developing an entirely new instrument. The decision was made, therefore, to adapt an existing assessment of critical thinking for use in the Rwandan context. As no cultural adaptation methods could be identified in the critical thinking literature, the authors worked together to develop a new methodology for accomplishing this objective.

Adaptation methodology

Although cultural adaptation of critical thinking assessments has not featured strongly in the field to date, there is a substantial literature regarding the selection of assessment tools for use in particular settings and circumstances (e.g. Ennis 2009). There are also international guidelines specifying best practice in developing tests for use in diverse cultural contexts (e.g. Hambleton 1994; Schmeiser and Welch 2006). Drawing on this literature, we articulated three specific research questions to guide our analysis of how critical thinking might be appropriately assessed in the Rwandan context:

1. Of the available tests of critical thinking, which format would be most appropriate for use in Rwanda?
2. How should the specific content of the chosen format be adapted, in order to ensure that the selected format provides a valid assessment of critical thinking in the Rwandan context?
3. Is the resulting adaptation valid for use in the study context?

These questions were answered via a mixed-methods strategy, comprising four stages: 1) selection of an appropriate test format; 2) adaptation of the test content; 3) field testing of the adapted content; and 4) full piloting of the final instrument.

Selection of an appropriate format

Ennis (2009) has argued that three steps must be completed when evaluating the validity of using a particular test of critical thinking in a particular context. First, one must make sure that the test is based on a "defensible conception of critical thinking and that the test does a reasonable job of covering that conception" (p. 75). Second, one must determine if the assessment content is appropriate for the situation (ibid.). Third, one must assess the validity of the scoring for the situation (ibid.). We applied Ennis' criteria to the list of existing assessments of critical thinking, in order to determine which format would be most appropriate for use in Rwanda.

Despite broad consensus around its importance, 'critical thinking' is one of the most contested constructs in education. Researchers and theorists argue vociferously over its definition, its relationship with similar constructs (such as 'reflective thinking' or 'problem solving') and its scope. One particularly contentious debate is between those who view critical thinking as a *generic* skill (e.g. Ennis 1985) and those who see it as *discipline-specific* (e.g. Moore 2004). There are also arguments around the role of dispositions in the definition of critical thinking, with some theorists conceptualising critical thinking as a skill or set of skills (i.e. Halpern 1996) and others suggesting that critical thinking also requires a dispositional element (i.e. Facione 1990).

The use of the term 'critical thinking' in the Rwandan higher education policy literature implies a generic conceptualisation of critical thinking, as there is clear assumption that critical thinking, although fostered within a particular academic discipline at university, should be applicable to a range of situations outside of the classroom. This conceptualisation resonates most closely with Kuhn's theory of critical thinking development (1999). Kuhn argues that

critical thinking requires cognitive skills (such as abstraction and the ability to differentiate theory from evidence), metacognitive strategies that allow for control over the thinking process, and a sophisticated epistemological worldview that recognises the uncertainty of knowledge. Kuhn's research has indicated that these "meta-knowing competencies" can be applied to ill-structured problems across domains, including in the 'real world', provided that both the ability and the disposition to utilise them are sufficiently developed through practice in individual content areas (Kuhn 1991, 1999, 2005). As references to 'critical thinking' in the Rwandan policy literature suggest that it is locally understood to be a generic construct comprising both ability and propensity elements, it was necessary to select an assessment format that could 'reasonably cover' this definition, in order to ensure the construct validity of the adapted instrument.

Given the size of the intended study sample, we considered only written formats for assessing critical thinking in the study population. Written critical thinking assessments fall into three broad categories: a) multiple-choice tests; b) essay-based tests; and c) performance-task-based tests (cf. Stein, Haynes and Unterstein 2003; Williams, Wise and West 2001). Of these three options, performance-task-based assessment maps most closely onto Kuhn's definition of critical thinking, as performance tasks are intended to simulate the "domain of real-world jobs suggested by activities found in … everyday practice" (Klein et al. 2007, 419). Although essay tests of critical thinking also allow for the consideration of evidence and the evaluation of arguments, the essay format does not offer the same authenticity as performance tasks. This undermines their effectiveness for assessing a respondent's ability and disposition to critically examine evidence when making a decision outside of a classroom setting. Performance-task-based assessment was, therefore, judged to be the most appropriate format for use in Rwanda, given the dominant conceptualisation of critical thinking in that context. Once the testing format was selected, there was little choice regarding the specific instrument, as, at the time of the study, the CLA, developed by the Council for Aid to Education in the U.S., was the only critical thinking test available which incorporated performance tasks. (The full CLA actually involves three components – two written essay sections and one performance task. However, as the essay sections are not relevant to the present study, only the performance task element is discussed here.)

Adaptation

Adaptation of content

CLA performance tasks require test takers to grapple with a fictional scenario, in which they must confront an 'ill-structured problem' in a workplace environment. They are given a number of documents in various formats to support them in their task, some of which are relevant to the task and some of which are not. They are then asked to explain how they would solve the challenge presented to them, based on the evidence provided.

Although this appeared to be an internationally applicable assessment format, we verified this assumption by presenting one of the CLA performance tasks – the 'Crime Reduction' Task, available in Chun (2008) – to a group of five Rwandan university students in March 2011. The focus group participants suggested that the assessment format should have both content and construct validity in Rwanda, as they agreed that a) students should develop the requisite skills to respond to such a task at university in Rwanda and b) university graduates in Rwanda would be expected to make similar decisions in the majority of workplace settings. However, they confirmed that the specific *content* of the CLA task would not be familiar to a Rwandan audience, as it focused on drug-related crime, an issue that is largely irrelevant in the Rwandan context.

As performance tasks are based on one overarching scenario, there is potential for differential performance depending on a given respondent's familiarity with the task content (Messick 1994). It was therefore desirable to create *two* adaptations of the CLA performance task, so that we could randomly administer the two versions in the study sample, thereby reducing the probability of any systematic interaction between task content and participant background. The two tasks created for the Rwanda study followed the Crime Reduction structure and format exactly but presented two locally relevant scenarios. In both versions, the respondent was asked to: assume the role of an intern in a government ministry; review the relevant evidence included in a number of documents; and outline the strengths and limitations of two policy options supposedly circulating in Parliament. As in the source version, each of the adapted tasks relied on seven documents and included three assessment questions. The first question asked respondents to evaluate the accuracy of a claim based on the evidence included in *one* document. The second question asked them to evaluate the strengths and weaknesses of one of the policy options, based on relevant information included in *all* of the documents. The final question asked respondents if they could make a decision between the two policies, based on all of the evidence in the documents. If they could make a decision, they were asked to explain which option they supported and why. If they could not, they were asked if they could

propose another solution, based on evidence in the documents. The tasks focused on contemporary problems in Rwanda: road accidents and malaria. As it was determined that the target population would have significant difficulties with a CLA-style online delivery platform, the adapted tasks were created as pen-and-paper assessments that could be administered in a group setting with invigilators available to answer questions and clarify instructions.

In addition to adapting the test content, we translated the tasks prior to piloting. There is substantial evidence to suggest that assessment in a foreign language underestimates the critical thinking ability of respondents, because respondents using a foreign language are required to devote a substantial proportion of their working memory to the comprehension of assessment questions (Floyd 2011; Just and Carpenter 1992; Takano and Noda 1993). In the Rwanda study, this situation could have been avoided by administering the assessment in Kinyarwanda, the language spoken by 99.4% of the Rwandan population (Samuelson and Freedman 2010). However, the authors' lack of fluency in Kinyarwanda prevented this possibility. Instead, we elected to write the assessment in both English and French (the two official languages of instruction in Rwanda, prior to 2010) and to allow participants to respond to the assessment questions in any language of their choosing (French, English or Kinyarwanda). The back-translation method, first advocated by Brislin (1970), was used to verify the conceptual equivalence of the French and English versions of the tasks.

Adaptation of scoring

We also suspected that we would need to adapt the CLA *scoring methodology*, as we had concerns about its validity in the Rwandan context. Although CLA scoring rubrics are not publicly available, the general methodology is discussed at length in CLA publications (e.g. Benjamin et al. 2009). The CLA scoring method relies on the assumption that complex cognitive tasks, such as performance tasks, require an integration of cognitive abilities that cannot be validly scored as individual skills (Klein et al. 2007; Shavelson 2010). In line with this general principle, the CLA does not assign individual scores for specific sub-skills (Benjamin, Chun, and Shavelson 2007). Instead, participants receive holistic scores for each of four assessed categories (Analytic Reasoning and Evaluation, Problem Solving, Writing Effectiveness, and Writing Mechanics). In order to ensure some reliability between scorers, CLA performance tasks are scored using a combination of analytic and holistic scoring (Benjamin et al. 2009). The analytic score is derived using a checklist of criteria. After assessing a given student's analytic score, a holistic score is assigned, indicating the

participant's overall ability in each of the general categories (ibid.). The final score is recalibrated to one overall score on a 1600-point scale that mirrors the Scholastic Aptitude Test (SAT), so that the SAT can be used as a control for incoming ability during analysis (CAE 2008). In addition to the lack of an SAT-type assessment in Rwanda (which limited the utility of a 1600-point scale), we had concerns about translating individual analytic scores into one overall score. During the development of the CLA, factor analysis was conducted on individual analytic scores, and, as all of the scores were found to load onto one underlying factor, it was deemed valid to use one overarching score to reflect participant ability (personal communication, May 22, 2012). However, we suspected that individual skills might load differently in the Rwandan context. As a result of these concerns, we elected to develop our own scoring method for use with the adapted assessment.

A rough scoring structure was determined during the development stage. As discussed above, the CLA is intended to assess respondent ability in four categories: Analytic Reasoning and Evaluation; Problem Solving; Writing Effectiveness; and Writing Mechanics. As we were less interested in participant ability to construct a well-written response, we elected to focus exclusively on the first two categories. The developers of the CLA define these categories as follows:

- Analytic Reasoning and Evaluation: "Interpreting, analysing and evaluating the quality of information; Identifying information that is relevant to a problem, highlighting connected and conflicting information, detecting flaws in logic and questionable assumptions; Explaining why information is credible, unreliable, or limited."
- Problem Solving: "Considering and weighing information from discrete sources to make decisions (draw a conclusion and/or propose a course of action) that logically follow from valid arguments, evidence and examples; Considering the implications of decisions and suggesting additional research when appropriate." (Chun 2008, 42)

The assessment of these individual skills relies on respondent use of the various documents included in the task, with each document playing a specific role. In the CLA performance task manual, each document type is accompanied by a description of the skills that respondents are expected to demonstrate as a result of analysing the document. Different documents may be expected to elicit a number of individual skills, including: determining that results from one study might not apply to a different setting; recognising possible sources of bias, or recognising the difference between correlation and causation (Chun 2008). As we intended to use the same document types as were used in the CLA tasks, we extracted a set of nine critical thinking skills from the CLA documentation that could be assessed via our adapted performance tasks:

1. The ability to recognise potential sources of personal bias (Skill A: Bias)
2. The ability to determine whether or not information is relevant to a situation (Skill B: Relevance)
3. The ability to recognise when a source of information is not credible or reliable (Skill C: Credibility)
4. The ability to identify statistical or methodological errors in presented information (Skill D: Errors)
5. The ability to determine whether or not information can be generalised and/or applied to other situations (Skill E: Generalisability)
6. The ability to recognise when there is a lack of information (Skill F: Missing Information)
7. The ability to evaluate whether or not information is connected and, if so, whether the data is conflicting or complementary (Skill G: Evaluation of Connections)
8. The ability to evaluate whether or not information supports or contradicts an argument (Skill H: Evaluation of Support)
9. The ability to draw on valid evidence when formulating a decision (Skill I: Use of Evidence)

These skills were then used as the basis of our adapted scoring method. We elected to use a criterion-referenced scoring methodology, in which each participant's response would be assessed against a pre-determined scale for each of the nine skills (Astin 1991). Draft scoring rubrics were therefore created to accompany each task, with each skill being assessed via a five-point scale (ranging from 1 [poor] to 5 [exemplary]). (Full scoring rubrics have not been included here due to space constraints but are available from the authors upon request.)

Field testing and piloting

In the final stage, a variety of qualitative and quantitative techniques were used to empirically test the validity of the adapted instrument, inspired by validation methods used to evaluate new CLA performance tasks (as discussed in Benjamin et al. 2009; CAE 2011; Chun 2008).

Validation methods

First, feedback was solicited from two 'expert' panels. The first panel, consisting of four Rwandan reviewers, reviewed the tasks for clarity, authenticity and any potentially offensive or alienating content. The second panel - comprising two American academics with expertise in critical thinking assessment - were asked to comment on the face validity of the adapted tasks and to judge the apparent comparability of the tasks.

Second, qualitative 'think aloud' techniques with Rwandan undergraduates were used to assess the content validity of the adapted tasks. Think aloud techniques are commonly used

when field testing performance assessments (Chi, Glaser, and Rees 1982), as such techniques allow developers to compare the skills 'demonstrated' by participants with the skills that a given assessment is intended to test (Lane and Stone 2006). Think aloud sessions are typically conducted individually, with one respondent sharing responses with a researcher (Forsyth and Lessler 1991). However, as CLA tasks are generally field tested through group think aloud sessions with university students (Shavelson 2010), a similar technique was used to field test the adapted tasks in Rwanda. In September 2011, each task was independently tested by a group of 3-5 Rwandan university students. The groups represented a range of disciplinary backgrounds and included both male and female participants. Each participant was given the option to either select his or her language of preference or to use both versions of the tasks simultaneously in order to aid in comprehension. Participants were first given half an hour to read the instructions and the content of one of the tasks. They were then asked to paraphrase the overall task scenario, re-articulate the assessment questions in their own words and discuss, as a group, how they would have responded to each question (as advocated by Jobe and Mingay 1989).

Following the think aloud sessions, a full pilot test was organised with 17 Rwandan students, representing both genders and a range of disciplinary backgrounds. In the pilot, the two tasks were distributed randomly within the group. Participants were allowed to choose between the French and the English version of their allocated task, although, in practice, the majority elected to use the English version. Two volunteers agreed to take both the French and English versions of their assigned task. As a result, 19 pilot responses were available for subsequent analysis. These responses were used to test the construct validity of the adapted tasks and the reliability of the scoring methodology.

The adapted tasks were then administered to a large random sample of 220 first- and fourth-year students attending three public universities in Rwanda. At each institution, four or five data collection sessions were organised, all of which were held in classrooms on the university campuses and invigilated by the lead researcher. The two task versions were randomly distributed within the sample (n=109 for Task 1; n=111 for Task 2). The sample was relatively evenly divided between first- and fourth-year students (57.3% first years versus 42.7% fourth years) and between Science and Social Science/Humanities subjects (58.6% and 41.4%, respectively). Most of the study participants were male (78.6% of the overall sample), due to the overrepresentation of male students at the three participating institutions, but the participants represented a broad range of backgrounds, in terms of secondary school type,

parental education level and socio-economic status. Scores from the administration of the adapted instrument within the full sample were used to test the comparability of the two task versions and to further evaluate the validity of the scoring method by investigating the underlying factor structure.

Finally, individual interviews were conducted with faculty members at two of the universities participating in the study. During the interviews, the faculty participants were asked to give their own definition of critical thinking and to share their perspectives on the relevance of the individual critical thinking skills referenced in the scoring rubric. In total, 18 interviews were conducted (9 at each institution). The participants represented a mix of gender, academic rank, and disciplinary background.

Results

Responses from the two expert panels gave initial support for the construct validity of the adapted tasks. The Rwandan panel confirmed that both tasks felt authentic and elicited skills that were likely to be fostered via a university education in Rwanda. The faculty panel were similarly positive in their evaluation of the face validity of the tasks, while also confirming that the two versions appeared to be parallel in structure and scoring.

The think aloud sessions confirmed that a performance task could be feasibly administered to university students in Rwanda, as none of the focus group participants had any difficulty understanding the instructions or re-articulating the content of the specific test questions. Furthermore, participant explanations of how they would respond to the tasks provided further support for the tasks' content validity, as their responses demonstrated the use of most of the individual skills included in the scoring rubric, e.g.:

> "I think here, the idea of [this M.P.] for promoting this mosquito net distribution? It was one way for her husband – her partner - to raise money." (demonstration of Skill A: Bias)

> "I have a doubt, I guess. … I'm more focused on how accurate this is. So, [this chart is from] Ethiopia and [the M.P] is using this chart as an example, but maybe there could be another study in Rwanda and then use that for the case of Rwanda. Of course, Ethiopia is … in East Africa, and you know, we can generalise, you know, to say this chart applies to Rwanda too, but what if Ethiopia … has a different level of malaria issues, you know? That's my concern." (demonstration of Skill B: Relevance and Skill E: Generalisability)

> "Maybe the problem which I have with this chart, ok, it is not showing which countries are… I think each country has its own maximum speed, and it's not showing which statistic are for what country" (demonstration of Skill F: Missing Information)

> "Yeah, based on these documents, they are strength for supporting the message of breathalyser tests. Like, from the Document A, it showed that the accident was caused by the drunk driver. And the Document B, they are showing that 60% … of road accidents were [because] they was drunk. And in the Document F also, they show that the people who drank frequently, they caused more accidents… Yeah, maybe there is a problem of, a limitation of this argument, according to the Document G where … there are some studies which are showing that you can underestimate the blood alcohol content [using a breathalyser] because of gender, body mass and physical exercise…. But, the overall, all documents are really showing the strengths of this method." (demonstration of Skill B: Relevance, Skill G: Evaluation of Connections, Skill H: Evaluation of Support, and Skill I: Use of Evidence)

A qualitative review of the written responses from the pilot phase further supported the content validity of the instrument, as all of the responses included evidence of at least some of the skills that the assessment was intended to evaluate. The pilot responses also supported the construct validity of the scoring method, as it was possible to score all of the responses using the rubrics, and the resulting scores were determined to be appropriate reflections of the ability demonstrated in each response. Evaluation of the pilot responses also confirmed the comparability of the individual scoring scales (i.e. the level of ability required for a 3 on one skill was found to be similar to the level of ability required for a 3 on another). However, it was not possible to test the validity of the full range of scores using the results from the pilot, as none of the pilot responses were judged to be of sufficient quality to merit a 4 or a 5 on some of the scales (e.g. Skill C: Credibility). As it was necessary to test the full range of scores prior to administering the instrument to a larger sample, a volunteer, purposively selected as likely to be an 'expert', was recruited to complete one of the tasks (as recommended by Baxter and Glaser 1998). Her responses were then scored against the rubrics in order to test the top range of each scale.

The pilot responses were also used to test the inter-rater reliability of the assessment scores. Three volunteer scorers were trained on the scoring methodology and then asked to score four pilot responses each (two of each task version). In total, six responses (three of each task version) were scored, with each individual response being evaluated by two of the three scorers. Inter-rater reliability of the scoring was tested through the calculation of intraclass correlation coefficients. Following the recommendation of Shrout and Fleiss (1979), a two-way random effects model was used, in which both the judges and the target ratings were treated as random effects. Analysis of each coefficient's difference from zero was assessed via an *F* test (see Table 1).

Table 1: Results of reliability analysis (n=6)

Skill	Intraclass Correlation Coefficient for Single Measures	95% Confidence Interval	Significance
A: Bias	.899	.649, .984	$F(5,10) = 27.727$ ($p<.001$)
B: Relevance	.714	.250, .948	$F(5,10) = 8.480$ ($p=.002$)
C: Credibility	.558	.041, .911	$F(5,10) = 4.784$ ($p=.017$)
D: Errors	.854	.529, .976	$F(5,10) = 18.483$ ($p<.001$)
E: Generalisability	.799	.406, .966	$F(5,10) = 12.932$ ($p<.001$)
F: Missing Information	.828	.468, .971	$F(5,10) = 15.400$ ($p<.001$)
G: Evaluation of Connections	.789	.386, .964	$F(5,10) = 12.229$ ($p=.001$)
H: Evaluation of Support	.700	-.090, .991	$F(2,4) = 8.000$ ($p=.04$)
I: Use of Evidence	.701	.230, .946	$F(5,10) = 8.024$ ($p=.003$)

The general rule of thumb is that tests with reliability coefficients of more than .7 can be used for individual-level analysis (Haertel 2006). Although the reliability coefficient for Skill C: Credibility was lower than this threshold, the overall results indicated adequate inter-rater reliability for the use of the scoring methodology, particularly given that inter-rater reliability coefficients for performance-task-based assessments, including the CLA itself, occasionally dip below .7 (Shavelson 2010).

The reliability of the adapted instrument was further assessed within the full sample by comparing the scores from the two task versions. It was not feasible to collect test-retest data from the same group of participants, given the duration of the assessment, but, as the groups of participants completing the two task versions were randomly selected from the same population, comparison of the scores obtained via the two tasks gave us a comparable measure of reliability. First, we investigated the absolute comparability of scores on the two tasks via a one-way MANOVA, with Task as the grouping variable. The analysis yielded significant results, using Pillai's trace $F(9, 188)=4.784$, $p<.001$. However, when Bonferroni's correction

was applied to adjust for the effect of running multiple tests (as suggested by Tolmie, Muijs, and McAteer 2011), a significant difference was only observed for two skills (Skill A: Bias and Skill D: Errors). The effect size for Skill A (assessed using partial eta squared) was modest (.100), while the effect size for Skill D was quite weak (.045). These results indicate broadly consistent measurement of critical thinking across the two task versions, with the two versions yielding a similar pattern of scores. The similarity in scores was further explored by investigating relative differences in the rank order profile of the nine individual skills. A rank-order correlation indicated a strong positive relationship between the scoring profile of the two versions [$r = .824$, p (one-tailed) $< .01$; $r_s = .795$, p (one-tailed) $< .01$], further confirming consistency in measurement of the underlying construct.

As discussed above, our final concern regarding the construct validity of our scoring method was whether it would be valid, in the Rwandan context, to implement the CLA methodology of combining the nine individual scores into one overarching score. We explored this question by conducting principal components analysis on the scores from the full sample, in order to investigate the underlying factor structure. The optimal solution for the analysis was determined using average variable complexity, a method that considers the average number of factors on which variables have an appreciable loading (i.e. >.3) and assumes that the solution with average variable complexity closest to one is the optimal solution (Tolmie et al. 2011). A seven-factor solution with varimax rotation was found to yield the least complexity, explaining 91% of the variance. Three of the skills (Skill G: Evaluation of Connections; Skill H: Evaluation of Support; and Skill I: Use of Evidence) were found to load on one latent factor, with the remaining six skills loading independently. Table 2 shows the factor loadings after rotation.

Table 2: Factor loadings after rotation

	Component						
Skill	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A: Bias							.973
B: Relevance		.903					
C: Credibility					.976		
D: Errors				.960			
E: Generalisability			.960				
F: Missing Information						.993	
G: Evaluation of Connections	.716	.463					
H: Evaluation of Support	.846						
I: Use of Evidence	.764						

These results present a different picture of the relationship between individual critical thinking skills from that observed during development of the CLA. Three of the skills (Skills G, H and I) are clearly correlated. In addition to loading onto one factor, combining these three skills generated an alpha coefficient of α=.739, indicating reliability of the resulting scale (Tolmie et al. 2011). This is theoretically unsurprising, as these skills - all of which would be classified as 'evaluation' skills (Kuhn 2005) - are likely to develop simultaneously. Individuals who have learned to recognise connections between pieces of evidence are also likely to be able to recognise how those pieces of evidence might connect with an argument and to use them when making an independent decision. In contrast, it is feasible to imagine that an individual could develop some of the additional six skills - all of which are better classified as 'inquiry' (rather than 'evaluation') skills (ibid.) - without developing the others (i.e. one could easily learn to assess the relevance of a piece of information without ever learning that sources differ in their credibility). One reasonable explanation for the relative independence of six of the skills, therefore, is that these skills might emerge independently as a result of explicit instruction or modelling.

In the U.S., most students are exposed to explicit modelling of both evaluation and inquiry skills in schools. However, evidence from the final validation method – the faculty interviews – suggests that this may not be the case in Rwanda. All of the faculty participants

indicated that critical thinking is viewed as an important component of a university education in Rwanda. However, when asked to define 'critical thinking', most described only the elements of the three 'evaluation' skills included in the scoring rubric. Of the 18 faculty participants, only five mentioned any of the other skills. If faculty members do not view 'inquiry' skills, such as questioning evidence, as a component of critical thinking, it is less likely that they will offer specific instruction in such skills or even model their use in the classroom. As evidence from studies in other contexts suggests that explicit modelling is crucial for the development of critical thinking skills (Marin and Halpern 2011), such a difference between the education systems could at least partially explain why some individual critical thinking skills, which are highly correlated in the U.S. context, may be relatively independent in the Rwandan student population.

Reflections on the adaptation process

The decision to complete a full adaptation process as part of the Rwanda study did add to both the study's cost and its timeline. The methods outlined in this paper spanned a period of approximately eighteen months, although much of this time was also dedicated to other aspects of the overarching study. In total, we estimate that the lead researcher devoted three months of full-time work to the adaptation methodology described in this paper. In terms of cost, the process required an additional trip from the U.K. to Rwanda, as well as some direct costs associated with the validation methods (i.e. minor incentives for participants and photocopying of the tasks for both field testing and piloting).

We also acknowledge that the adaptation process would have been more challenging in an unfamiliar context. As the lead researcher had lived and worked in Rwanda for a number of years prior to commencing the study, she had pre-existing relationships with a number of Rwandan volunteers who were happy to assist with the expert review process, the field testing and the piloting. Her knowledge of the study context also allowed her to personally complete the initial adaptation of the tasks. When working in an unfamiliar context, a similar strategy would only be possible with the expertise and participation of local researchers.

However, despite these costs and potential challenges, our experience strongly recommends the use of a cultural adaptation method when seeking to assess critical thinking in a new cultural context. If the original version of the CLA had been used in the Rwanda study, it is clear the validity of the study results would have suffered substantially. First, the unfamiliar content of the CLA performance tasks would likely have introduced a significant amount of

construct-irrelevant variance into the scoring distribution. The final analysis in the Rwanda study ultimately indicated that Rwandan students are *not* improving substantially in their critical thinking ability during their time at university (Schendel 2015). As this finding raised concerns for many within the Rwandan higher education sector, it is likely that the use of a foreign assessment would have generated understandable scepticism around the study results. In contrast, our use of an explicitly adapted instrument supported local stakeholder confidence in the study results. Second, our principal components analysis results suggest that it would not have been valid to use the original CLA scoring methodology in the study, given differences in the way that individual critical thinking skills appear to interrelate in the Rwandan context. Our decision to verify the factor structure underlying the scoring method had particular implications for our study design, as our initial intention had been to combine the individual skill scores into one overarching score in order to use regression during analysis. Following the principal components analysis, we revised this analytical strategy, electing instead to retain the independence of the individual scores. Finally, the adaptation method exposed the need for a number of changes to the assessment tool that would not otherwise have been obvious. For instance, during the think aloud sessions, participants raised concerns about the order of materials presented in the test booklet. In the original CLA, the test questions appear before the individual documents. However, the focus group participants suggested that Rwandan students would be likely to try to answer the questions as soon as they saw them, without realising that they should read the supporting documentation first. They therefore proposed moving the question page to the end, so that participants would have to read both the scenario and the documents before reaching the questions. If this potential pitfall had not been identified at the outset of the study, the validity of the test results would have been severely compromised.

Conclusion

It has long been acknowledged that, the more an assessment has "indigenous validity" (Irwin, Klein, and Townsend 1982) within a population, the more likely it is that the assessment will actually sample from the intended domain under consideration (Cole and Scribner 1974). The validity of any study of learning outcomes is, therefore, significantly improved if it relies on an assessment tool that has been created for or adapted to the particular study context. Although this is best accomplished through the development of local instruments and methods, local development of complex assessments of cognitive ability is not

a realistic option in many contexts. When faced with this dilemma, many research teams elect to simply translate the content of assessments imported from other contexts. However, direct translation is not sufficient for assessments of higher-order cognitive skills, as attention must also be paid to the familiarity of the assessment structure and the validity of the scoring method for the local context. In contrast, the use of an in-depth cultural adaptation methodology like the one outlined in this paper can help to structure the adaptation process and evaluate the validity and reliability of the resulting tool. As the entire selection, adaptation and evaluation process presented here was completed within three months, with only one primary researcher and very limited research funding, our experience demonstrates that cultural adaptation is not only desirable but feasible, even for small research teams.

Although designed for use in Rwanda, the adaptation methodology presented here could be successfully implemented elsewhere, including in other contexts similarly constrained by limited resources. This, in turn, would be a substantial contribution to the field, as the development of other culturally relevant assessments could generate a wealth of culturally sensitive empirical data, which could then be leveraged by both universities and international organisations to improve academic quality at higher education institutions across the Global South.

Acknowledgements

The authors would like to acknowledge the pilot participants who volunteered their time to help us with this research.

References

Arum, R., and J. Roksa. 2011. *Academically adrift : limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Ashcroft, K. and P. Rayner. 2011. *Higher Education in Development: Lessons from sub-Saharan Africa*. Charlotte, NC: Information Age Publishing, Inc.

Astin, A. W. 1991. *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education*. New York, NY: American Council on Education and Macmillan Publishing Company.

Bataineh, R. F., and L. H. Zghoul. 2006. "Jordanian TEFL Graduate Students' Use of Critical Thinking Skills (as Measured by the Cornell Critical Thinking Test, Level Z)." *International Journal of Bilingual Education and Bilingualism* 9 (1): 33-50.

Baxter, G. P., and R. Glaser. 1998. "Investigating the Cognitive Complexity of Science Assessments." *Educational Measurement: Issues and Practice* 17 (3): 37-45.

Benjamin, R., M. Chun, C. Hardison, E. Hong, C. Jackson, H. Kugelmass, A. Nemeth, and R. Shavelson. 2009. *Returning to Learning in an Age of Assessment: Introducing the*

Rationale of the Collegiate Learning Assessment. New York, NY: Council for Aid to Education.

Benjamin, R., M. Chun, and R. Shavelson. 2007. *White Paper. Holistic Tests in a Sub-Score World: The Diagnostic Logic of the College Learning Assessment*. New York, NY: Council for Aid to Education.

Blaich, C., and K. Wise. 2010. *Wabash National Study of Liberal Arts Education 2006-2009: Overview of Findings from the First Year*. Wabash College Center of Inquiry. Available from http://www.liberalarts.wabash.edu/study-4th-year-data/.

Boni, A., and D. Gasper. 2012. "Rethinking the quality of universities: How can human development thinking contribute?" *Journal of Human Development and Capabilities* 13 (3): 451-470.

Brislin, R. W. 1970. "Back-translation for cross-cultural research." *Journal of Cross-Cultural Psychology* 1 (3): 186-216.

CAE (Council for Aid to Education). 2008. *CLA Frequently Asked Technical Questions*. New York, NY.

CAE (Council for Aid to Education). 2011. *Architecture of the CLA Tasks*. New York, NY.

Chi, M. T. H., R. Glaser, and E. Rees. 1982. "Expertise in Problem Solving." In *Advances in the psychology of human intelligence*, edited by R. J. Sternberg. Hillsdale, NJ: Erlbaum.

Chun, M. 2008. *Introduction to Performance Tasks*, *CLA in the Classroom*. New York, NY: Council for Aid to Education.

Cole, M., and S. Scribner. 1974. *Culture and thought: a psychological introduction*. New York, NY: John Wiley.

Davies, M., and R. Barnett. 2015. "Introduction". In *The Palgrave Handbook of Critical Thinking in Higher Education*, edited by M. Davies and R. Barnett. New York, NY: Palgrave MacMillan.

Ennis, R. H. 1985. "A Logical Basis for Measuring Critical Thinking." *Educational Leadership* 43 (2): 45-48.

Ennis, R. H. 2009. "Investigating and Assessing Multiple-Choice Critical Thinking Tests." In *Critical Thinking Education and Assessment: Can Higher Order Thinking Be Tested?*, edited by J. Sobocan and L. Groarke. London, Ont.: Althouse Press.

Facione, P. A. 1990. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Executive Summary*. Millbrae, CA: California Academic Press.

Floyd, C. B. 2011. "Critical thinking in a second language." *Higher Education Research and Development* 30 (3): 289-302.

Forsyth, B. H., and J. T. Lessler. 1991. "Cognitive laboratory methods: a taxonomy." In *Measurement errors in surveys*, edited by P. P. Biemer, S. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudan. New York, NY: John Wiley.

Haertel, E. H. 2006. "Reliability." In *Educational Measurement*, edited by R. L. Brennan. Westport, CT: American Council on Education and Praeger Publishers.

Halpern, D. F. 1996. *Thought and knowledge : an introduction to critical thinking*. (3rd Ed.). Mahwah, N.J: L. Erlbaum Associates.

Hambleton, R. K. 1994. "Guidelines for Adapting Educational and Psychological Tests: A Progress Report." *European Journal of Psychological Assessment* 10 (3): 229-244.

Irwin, M., R. E. Klein, and J. W. Townsend. 1982. "Indigenous versus Construct Validity in Cross-Cultural Research." In *Cross-Cultural Research at Issue*, edited by L. L. Adler. New York: Academic Press, Inc.

Jobe, J. B., and D. J. Mingay. 1989. "Cognitive research improves questionnaires." *American Journal of Public Health* 79 (8): 1053-1055.

Just, M. A., and P. A. Carpenter. 1992. "A Capacity Theory of Comprehension: Individual Differences in Working Memory." *Psychological Review* 99 (1): 122-149.

Klein, S., R. Benjamin, R. Shavelson, and R. Bolus. 2007. "The Collegiate Learning Assessment: Facts and Fantasies." *Evaluation Review* 31: 415-439.

Kuhn, D. 1991. *The skills of argument*. Cambridge: Cambridge University Press.

Kuhn, D. 1999. "A Developmental Model of Critical Thinking." *Educational Researcher* 28 (2): 16-25+46.

Kuhn, D. 2005. *Education for Thinking*. Cambridge, MA: Harvard University Press.

Lane, S., and C. A. Stone. 2006. "Performance Assessment." In *Educational Measurement*, edited by R. L. Brennan. Westport, CT: American Council on Education and Praeger Publishers.

Lombard, K., and M. Grosser. 2004. "Critical thinking abilities among prospective educators: ideals versus realities." *South African Journal of Education* 24 (3): 212-216.

Lombard, K., and M. Grosser. 2008. "Critical thinking: are the ideals of OBE failing us or are we failing the ideals of OBE?" *South African Journal of Education* 28: 561-579.

Luescher-Mamashela. T. M., V. Ssembatya, E. Brooks, R. S. Lange, T. Mugume and S. Richmond. 2015. "Student Engagement and Citizenship Competences in African Universities." In *Knowledge Production and Contradictory Functions in African Higher Education*, edited by Nico Cloete, Peter Maassen & Tracy Bailey. Cape Town: African Minds.

Lun, V. M.-C., R. Fischer, and C. Ward. 2010. "Exploring cultural differences in critical thinking: Is it about my thinking style or the language I speak?" *Learning and Individual Differences* 20 (6): 604-616.

Marin, L. M. and D. F. Halpern. 2011. "Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains." *Thinking Skills and Creativity* 6: 1-13.

Mayer, R. E., and M. C. Wittrock. 1996. "Problem-solving transfer." In *Handbook of educational psychology*, edited by D. C. Berliner and R. C. Calfee. New York, NY: Prentice Hall International.

Messick, S. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23 (2):13-23.

Moore, T. 2004. "The Critical Thinking Debate: How general are general thinking skills?" *Higher Education Research and Development* 23 (1): 3-18.

Obura, A. P., UNESCO, and International Institute for Educational Planning. 2003. *Never again : educational reconstruction in Rwanda*, *Education in emergencies and reconstruction. Case studies*. Paris: International Institute for Educational Planning.

Osman, R.M., and B.N. Githua. 2009. "Analysis of Students' Critical Thinking Skills by Location, Gender and the Skills Relationship to Motivation to Learn Academic Disciplines in Amoud University, Somaliland." *Journal of Technology and Education in Nigeria* 14 (1 & 2): 1-10.

Phan, H. P. 2011. "Deep Processing Strategies and Critical Thinking: Developmental Trajectories Using Latent Growth Analyses." *The Journal of Educational Research* 104 (4): 283-294.

Rieckmann, M. 2012. "Future-oriented higher education: Which key competencies should be fostered through university teaching and learning?" *Futures* 44 (2): 127-135.

Rogoff, B. 2003. *The cultural nature of human development*. Oxford: Oxford University Press.

Saavedra, A. R., and J. E. Saavedra. 2011. "Do colleges cultivate critical thinking, problem solving, writing and interpersonal skills?" *Economics of Education Review* 30: 1516-1526.

Samuelson, B. L., and S. W. Freedman. 2010. "Language policy, multilingual education, and power in Rwanda." *Language Policy* 9: 191-215.

Schendel, R. 2015. "Critical thinking at Rwanda's public universities: Emerging evidence of a crucial development priority." *International Journal of Educational Development* 42: 96-105.

Schmeiser, C. B., and C. J. Welch. 2006. "Test Development." In *Educational Measurement*, edited by R. L. Brennan. Westport, CT: American Council on Education and Praeger Publishers.

Serpell, R. 2007. "Bridging between orthodox western higher educational practices and an African sociocultural context." *Comparative Education* 43 (1):23-51.

Shavelson, R. 2010. *Measuring College Learning Responsibly: Accountability in a New Era*. Stanford, CA: Stanford University Press.

Shrout, P. E., and J. L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86 (2): 420-428.

Stein, B. S., A. F. Haynes and J. Unterstein. 2003. *Assessing Critical Thinking Skills*. Paper presented at the SACS/COC Annual Meeting. Nashville, TN.

Takano, Y., and A. Noda. 1993. "A Temporary Decline of Thinking Ability during Foreign Language Processing." *Journal of Cross-Cultural Psychology* 24 (4): 445-462.

Tolmie, A., D. Muijs, and E. McAteer. 2011. *Quantitative methods in educational and social research using SPSS*. Maidenhead: Open University Press.

United Nations. 2015. *Sustainable Development Goals*. Available at: www.un.org/sustainabledevelopment/sustainable-development-goals/.

UNESCO. 2009. *Final Report: World Conference on Higher Education*. Paris: UNESCO.

Williams, K., S. L. Wise and R. F. West. 2001. *Multifaceted Measurement of Critical Thinking Skills in College Students*. Paper presented at the Annual Meeting of the American Educational Research Association. Seattle, WA.