Examining the Validity and Reliability of the CALL Formative Leadership Assessment: Pilot Study Results

Eric Camburn & Jason Salisbury
University of Wisconsin-Madison

This paper was prepared for presentation at the 2012 meeting of the American Educational Research Association. The research reported in this paper was supported by a grant from the U.S. Department of Education, Institute for Educational Sciences. The opinions expressed are solely the responsibility of the authors. Please direct correspondence to Eric Camburn, camburn@wisc.edu.

Background

There is a growing need for schools to have access to high quality evidence about conditions that support student achievement. The recent arrival of formative tools that assess the work of school leaders to advance student learning directly addresses this need. In order for evidence from these tools to be useful and trustworthy for practitioners, the validity and reliability of the evidence they produce must be established. This paper takes a step in that direction by presenting efforts undertaken to assess and improve the validity of the Comprehensive Assessment of Leadership for Learning (CALL) survey. We begin by describing the survey instrument. We then discuss a series of strategies that have been employed, including focus groups, reviews of survey experts, and a pilot study, that attempt to shed light on and improve survey validity.

The CALL Survey

The CALL tool captures evidence of school leadership practices via a web survey that is completed by teachers and leaders within a school. The tool measures leadership practice in five domains which themselves are measured with multiple subdomain scales.

When first accessing the survey, respondents are asked to complete a user identification section. Information collected at this stage includes: the name of the respondent's school and school district and respondent role (teacher, school leader, district leader, other). This information was collected through drop-down menus containing pre-loaded response options. In this initial section of the survey respondents were also asked to create a log-in and password that allowed participants to exit and re-enter a saved survey. During the pilot study approximately 75

individuals did not provide a school name thus rendering their data unusable. In subsequent revisions of the instrument, steps have been implemented to reduce this kind of missing data.

Questions were presented either as individual items or in tables combining multiple items. Individual item questions were designed with one question per page and asked users to hit a 'next' button to move to the next question. These questions were presented with a stem or construct that provided the user with a heading indicating the topic of the question such as "Assessment of teaching practices in our school." The topic was followed by a question, such as "How often do teachers in your school receive formal evaluations?" Following the question users were presented with 3-5 response options with radial buttons for answer selection.

The second type of question on the survey was table format; tables were used to capture multiple items in a single question and usually dealt with frequency of practice. As with single item questions, tables began with a construct and then had 3-5 questions embedded in the tables. As with single item questions, there was only one table per page and users needed to hit a 'next' tab to move onto the next question.

User responses were automatically saved once they hit the 'next' button and moved onto the next question. However, there was a 'previous' button that allowed users to move backward in the survey. If a user moved backward in the survey their previous response was still shown, but they could change their response and hit 'next' to save their updated response. Users also had the option of hitting a 'save and exit' button in each screen if they wanted to save a response and exit the survey. Lastly, there was a status bar on the top of the user's screen indicating the percentage of the survey they had completed.

As part of the pilot study we conducted a rigorous quality control assessment of the instrument which tested the accuracy with which the instrument captured participants' responses

and the accuracy of data export procedures. This assessment resulted in data capture and export functionality that was verified to be 100% accurate.

Survey Validation

The CALL research team has undertaken a rigorous validation of the CALL instrument. Validation included both qualitative and quantitative analyses of survey content, structure, and reliability. Qualitative analyses included focus groups focusing on question content and wording, interviews of pilot participants focusing on survey design and content, and a quality control analysis ensuring data entered into the survey matched data retrieved from the instrument. Quantitative analyses included various psychometric analyses and a variance decomposition analysis. Through these targeted analyses the CALL team has been able to create an updated version of the CALL instrument to be piloted in the spring and fall of 2012. The sections below provide a detailed description of the individual validation stages for the CALL instrument.

Focus Groups

Over a two-month period in the spring of 2010 a series of focus groups were conducted with two groups of practitioners. The purpose of the focus groups was to obtain practitioner input on survey content and question wording. Each focus group, one from a middle school and one from a high school, was made up of approximately 7 practitioners. Each focus group met 7 times and in the final focus group meeting both middle and high schools groups met together. Participants received a stipend of $1,000 and complimentary dinners during weekly meetings. Also, in each focus group two members of the research team were present, one serving as a facilitator and the other as note taker.

The first of the 7 focus group meetings was used to introduce participants to the goals of the CALL project, highlight expectations for focus group participants, and review a draft survey.

Group members were provided with information about the five domains making up the CALL instrument, the research supporting those domains, the historical context of the tool's development, and the formative feedback nature of the instrument. Once the groups were versed in the background of CALL they were provided a paper version of the survey to complete and record notes on. Participants were asked to take notes related to their general impression of the survey or to write about other things that peaked their interest. After reading the instrument and taking notes, focus group participants shared their general thoughts of the survey as well as their general concerns. Participant comments were recorded by the research team through field notes and by collecting the copies on which participants had recorded their notes. Generally, participant comments during the session were centered on the structure and content of the survey as a whole as opposed to individual items.

The next five sessions were each devoted to one of the domains measured on the survey. All five sessions followed the same format; focus group members were provided an overview of the domain and then given approximately 20 minutes to complete a paper copy of the survey items in that domain. While taking the survey participants were encouraged to take notes on anything that caught their attention (question content, missing material, subjective wording, double barreled questions, etc.). Upon completion of the paper survey the facilitator went through the domain item by item at which point participants were asked to voice concerns or comments on the given item during the review of items the facilitator projected a word document of the survey using an LCD projector. Typically multiple individuals had comments about items and dialogue would begin about item in question. During this process researchers would ask clarifying questions and record conversations in the form of field notes. Once the group had reached consensus on the given question the facilitator would use track changes to document the

edits the focus group had suggested. This process was repeated until all items of interest had been discussed. The final step during these five meetings was to cover items that focus group felt should be added to the domain in order to accurately capture the practices of a school related to specific construct.

Following the middle and high school focus group sessions the CALL research team would meet to share the results of the respective groups. During these sessions the team would look at the comments from the two focus groups and work to make changes to the instrument that aligned with both sets of comments. Typically there was agreement between the high school and middle school groups, but in cases where there was disagreement the team would discuss the difference and apply their professional judgment to make revise the instrument to best address both groups' concerns.

The final focus group meeting brought the middle and high school groups together. Participants were provided with an updated version of the survey to complete and take notes on. Following this exercise the group shared out and discussed any lingering concerns over item construction, item content, or missing items. As with previous focus group sessions the research team recorded participant comments and collected the paper copies of the instrument as another form of data. During the last part of the final focus group participants were shown screen shots of the online version of the CALL survey and asked for comments or suggestions. As a whole the focus group was pleased with the online formatting of the instrument and had no major suggestions for improvement.

Following the final practitioner focus group the research team met to discuss the suggested changes made by participants and come to consensus on possible edits. The result of

these focus groups and synthesizing meetings was a version of the instrument ready to be converted into an online tool for pilot schools to take.

Survey Expert Instrument Review

In January of 2012 CALL contracted with the University of Wisconsin Survey Center (UWSC) to provide feedback on updated survey questions, design a new web-based CALL instrument, provide data management services for CALL data, and advise the CALL team in the second round of piloting CALL. One of the largest university-based survey centers in the country, the center conducts thousands of telephone, face-to-face, mail, and web surveys and focus groups annually and achieves consistently high response rates across all survey methodologies. The UWSC has particularly strong expertise in survey design and question wording under the direction of Nora Cate Schaeffer, a nationally recognized expert in these areas.

UWSC questionnaire experts reviewed the CALL survey and made detailed suggestions for improvements based on current research in the area of survey development. These suggested changes included using common language that all participants will understand, eliminating mouse-over definitions for technical terms in favor of definitions embedded in the stems of questions, adding additional response options to eliminate double-barreled responses, and ensuring that all questions are worded as questions as opposed to statements. Additionally, the UWSC recommended the addition of screener questions to allow participants to skip questions that are not relevant to them or their schools.

Pilot Study

An initial online version of CALL the instrument was piloted in over 70 schools during the winter of 2011. These pilot schools were broken into two groups; the first were schools that

were all located in Mississippi who were solicited to take both CALL and VAL-ED[1]. In total, over 65 schools were included in this section of the pilot study. The second group included six schools from two different districts in Wisconsin that were in close geographic proximity to the research team. Schools in this second group were not asked to take the VAL-Ed leadership assessment, but were instead asked to participate in exit style interviews following the completion of the CALL instrument. By the end of the initial pilot of CALL over 1,700 teachers and over 150 school leaders had taken the instrument in two states.

Mississippi Pilot Schools

A convenience sample of pilot schools were recruited in Mississippi through a project conducted by the American Institute for Research[2] (AIR). The AIR project asked schools to take the Working Conditions survey and Val-Ed. Ultimately, 1,600 teachers and over 65 school leaders in 65 Mississippi schools participated in the pilot. This large, diverse sample of schools provided valuable insight into the survey's reliability and validity, and into the degree to which practitioners' perceptions of school leadership tended to diverge or converge within schools. School leaders were provided a link to the instrument by AIR in the spring of 2011 and given a two-week window in which to have staff complete the survey. The research team at AIR completed all initial and follow-up communication with Mississippi schools.

Wisconsin Pilot Schools

For the Wisconsin component of the pilot, the research team intentionally selected two nearby districts. The rationale behind their inclusion in the study was two-fold; first they were recruited due to previous relationships with the research team. This decision was made to

[1] VAL-ED is 360^0 leadership assessment developed by a research team and Vanderbilt University. For more information see http://www.valed.com

[2] This partnership was started with Learning Point Associates (LPA), but LPA became part of the American Institute for Research during the piloting of CALL.

increase access to the schools and to allow the maximum number of exit interviews possible. The second rationale was the districts diverse contexts; we believed collecting rich qualitative data from two contrasting districts would increase the variation in pilot user responses to the CALL instrument. District A is located in an urban setting and faces many of the traditional challenges facing urban districts in the United States. District B is a wealthier district that is a hybrid of a rural and suburban context. See Table 1 for a description of the two districts.

Table 1
Description of Wisconsin School Districts

District	# of Schools	# of participant schools	# of staff	# of teachers	# of participant staff	# of students in district	% Asian	% Black	% Latino/a	% Native American	% White	% ELL	% FRL
A	5	2	170	90	50	1,400	0.7	1.4	7.2	0.1	89.4	2.8	25.1
B	36	4	2,580	1,460	192	21,100	1.5	26.8	24.1	0.4	46.0	13.4	59.4

Wisconsin schools were recruited through informational meetings held at each school. During these meetings we presented the school-level and district-level leadership with information about CALL. Leadership teams were encouraged to ask questions during these meetings. Beyond information being shared about the CALL instrument and possible benefits for individual schools, our research team and the schools identified point people at each school and planned how to best implement the survey in each school.

There was some variation in the way Wisconsin pilot schools took the CALL survey. One high school used a staff-development day to have all teachers complete the survey at once. During this whole-staff activity a representative from the CALL team was present at the school to share information about the instrument, answer any questions, and help trouble shoot potential issues. We are aware of how the presence of the principal as the facilitator could influence

teacher responses, but are confident this did not happen based on the anonymity of the survey. The other five Wisconsin schools took the survey similarly to Mississippi schools. The schools were provided a link to access the survey and asked to distribute the link to their staff. Follow-up communications regarding survey completion was left to the leadership team at respective schools.

Once the Wisconsin schools completed taking the CALL instrument a qualitative case study was completed for each of the six schools. To complete this case study we conducted three rounds of interviews. The first round of interviews was with building principal and related to building specific practices and initiatives that should appear in the CALL data. During round two of interview we communicated with teachers about their experiences in taking the survey. This round of interviews specifically investigated if individuals understood the intended meaning of questions and their overall thought process while completing the instrument. For the final round of interviews we spoke with building principals about the feedback system of the survey and final thoughts on the process of participating in CALL. By generating a case of current practices in each of the schools we were able to capture the accuracy with which our instrument captured those practices. Additionally, interview questions about individual's experiences taking the instrument provided data centered on if participants understood questions as they were designed and if participants felt the wording of the questions was generous enough to capture practices that may have site-specific terminology associated with them.

Preliminary Pilot Results

The CALL survey was designed to create summary measures of leadership in 5 domains which are further divided into 21 sub-domains. Every item in the CALL survey is designed to measure an aspect of school leadership in a specific sub-domain. Practices within the same sub-

domain are believed to co-occur more commonly that practices in different sub-domains and practitioners who engage in a particular practice are believed to be more likely to engage in other practices within the same subdomain. In light of this reasoning, statistically, we expect practitioners' scores to be higher when they engage in more practices within a subdomain and lower when they engage in fewer practices. We also interpret strong intercorrelations among items intended to measure the same sub-domain as evidence that the items are measuring a common, underlying sub-domain.

We conducted three sets of statistical analyses to test these ideas.

Reliability analyses

We first conducted reliability analyses as an exploratory step to check the degree of dimensionality among items intended to measure the same subdomain. Table 2 contains Cronbach's Alpha statistic for each sub-domain scale. Reliability is a basic measure of the validity of a scale. Conceptually, reliability is defined as the degree to which a scale is free from errors of measurement. Measurement errors will be higher to the extent that different measurements of the same person vary. For example, if we give Jill a math test on Tuesday and the same test on Friday, we would expect fairly similar results. If we do this for all of the children in a school and get very different results between Tuesday and Friday our test is unreliable. Reliability is operationalized as a measure of the degree of consistency between multiple, equivalent measurements of the same construct. Reliability is higher when multiple measurements are more consistent with each other and lower when measurements are less consistent. An important property of reliability statistics is that they tend to increase with a greater number of measurements. To extend this example to the CALL data, multiple survey items that measure a particular sub-domain can be viewed as multiple measurements of a

construct. For example, the CALL survey includes 6 items that are intended to measure sub-domain 1.2 "Formal Leaders are Recognized as Instructional Leaders."

Our goal in instrument design was to achieve a reliability of at least .7 for each of the sub-domains. Reliability analysis based on the CALL pilot survey provided mixed results in achievement of that goal, with initial Chronbach's Alpha reliability scores of .7 or above for 11 of the 20 sub-domain scales. For each scale with a reliability score below .7, we have reviewed items in that scale and have added items, or revised items to improve reliability. The reliability analysis will be repeated for CALL Version 2.0 following administration of the revised survey this Spring.

Table 2: Cronbach's Alpha Reliability Coefficients by Sub-Domain

Subdomain	**Cronbach's Alpha Reliability**	**Number of Items**
1.1 Maintaining a school-wide focus on learning	.717	5
1.2 Formal Leaders are Recognized as Instructional Leaders	.799	6
1.3 Collaborative design of integrated learning plan	.612	3
1.4 Providing appropriate services for students who traditionally struggle	.176	4
2.1 Formative assessment of student learning	.674	4
2.2 Summative Evaluation of Student Learning	.572	4
2.3 Formative Evaluation of Teaching	.741	4
2.4 Summative Evaluation of Teaching	.774	6
3.1 Collaborative school-wide focus on problems of teaching and learning	.806	6
3.2 Professional learning	.521	5
3.3 Socially distributed leadership	.841	12
3.4 Coaching & Mentoring	.847	8
4.1 Personnel Practices	.317	5
4.2 Structuring & Maintaining Time	.512	3
4.3 School Resources Focused on Student Learning	.780	6

4.4 Integrating External Expertise into School Instructional Program	.536	4
4.5 Coordinating and Supervising Relations with Families & External Communities	.763	6
5.1 Clear, Consistent, & Enforced Expectations for Student Behavior	.720	5
5.2 Safe Learning Environment	.585	3
5.3 Student Support Services Provide Safe Haven for Students Who Traditionally Struggle	.870	12

Rasch Analyses

We attempted a more robust examination of dimensionality among items intended to measure the same subdomain by conducting a series of analyses which fit CALL pilot data to the Rasch item response theory model. The CALL survey was designed with the intention of measuring specific sub-domain constructs. Unlike exploratory factor analysis, the Rasch model is not designed to identify multiple constructs. Instead, the Rasch model assumes a set of items subjected to analysis is intended to measure a single construct.

Rasch analyses produce two item fit statistics called "Infit Mean Square" and "Outfit Mean Square" which indicate the degree to which responses to particular items depart from model assumptions. To understand the meaning of these statistics it is useful to have a rudimentary understanding of a number of key assumptions of the model. The model produces a "difficulty" statistic for each item. With Likert type survey questions, items that are more strongly endorsed (e.g. people choose 4s and 5s on a rating scale ranging from 1-5) by the *most* people have the *lowest* "difficulty." Items that are strongly endorsed by the *fewest* people have the *highest* "difficulty." The rank order of item difficulties represents the scale against which people are measured, and for data to fit the model well, peoples' responses to items across the difficulty hierarchy should be consistent with their scale score. Specifically, the model fits well when people at the low end of the scale do not strongly endorse the most difficult items (i.e. at

the upper end of the scale), and when people at the high end of the scale *do* strongly endorse the most difficult items. Departures from this pattern indicate model misfit, and Infit and Outfit mean square statistics indicate the degree to which such misfit is occurring. A value of 1 for these statistics indicates responses to a particular item fit the model perfectly. Values lower and higher than 1 indicate Reponses to a particular item misfit the model.

The two item fit statistics can be viewed as indicative of multidimensionality in items intended to measure a sub-domain. In other words, item misfit is indicative that an item may not be serving as good indicator of the intended sub-domain and may therefore be an indicator of some other construct. Our analyses flagged such misfitting items as candidates for item revision or deletion. For this analysis, we were concerned with Infit and Outfit statistics substantially greater than 1, and we used a threshold of 1.2 to flag items with particularly high levels of misfit. While Infit statistics indicate the degree to which responses to an item depart from the expected difficulty hierarchy, Outfit statistics are sensitive to unexpected outlier responses. Table 3 displays the number of items per sub-domain scale with Infit and Outfit mean square statistics that exceed 1.2.

Table 3: Rasch Item Fit Statistics for Sub-Domain Scales

Sub-domain	Number of Items	Items with Infit Mean Square > 1.2	Items with Outfit Mean Square > 1.2
1.1	5	1	2
1.2	6	0	0
1.3	3	0	0
1.4	7	1	2
2.1	8	2	2
2.2	4	0	0
2.3	4	0	0
2.4	7	1	1
3.1	6	2	2
3.2	5	1	1

3.3	12	4	4
3.4	10	2	3
4.1	6	1	1
4.2	3	0	0
4.3	6	1	1
4.4	4	0	0
4.5	5	1	1
5.1	5	1	1
5.2	3	0	1
5.3	12	1	0

The results in Table 3 indicate a fairly modest degree of item misfit. Generally speaking, we see considerable evidence that the clusters of items intended to measure CALL sub-domains are tapping a single underlying construct. Six of the sub-domain scales did not contain any items whose infit statistics exceeded 1.2. An additional 8 scales contained only 1 item whose infit statistics exceeded that threshold. There were of course exceptions to this general pattern, and the analyses identified a sub-domains (3.3 and 3.4) in which more than 2 items had exceptionally high item misfit. The results of this analysis were used to identify items for further scrutiny.

Variance Decomposition Analysis

We conducted third set of analyses that examined levels of within school agreement in CALL sub-domain scores and the reliability of school level means. We felt that gaining an understanding of the validity of school-level sub-domain scales was important since the CALL tool is designed to provide formative feedback at the school level. We further believed that evidence about the degree of agreement (or lack thereof) among staff within a school could also be an instructive piece of information for schools.

To investigate these issues we fit a series of two-level hierarchical linear models (HLM) which nested teachers within schools. The level 1 (teacher) equation modeled teachers' sub-

domain scores as a function of a school mean and residual error representing variation unique to each teacher. At level 2, school mean sub-domain scores were modeled as a function of an overall mean and an error term representing unique school effects. In addition to producing estimates of overall sub-domain means, these models also separated the total variation in the sub-domain scores into variation lying within-schools and between-schools. Finally, the HLM models produced estimates of the reliability of teacher scores and school means. The results of the HLM models are summarized in Table 4.

We found that CALL sub-domain scores are able to discriminate between schools quite reliably. Of the 20 sub-domains, all but three had school level reliabilities exceeding .70. This suggests that the CALL survey can reasonably accomplish one of its most basic purposes—distinguishing nature and prevalence of leadership practices in a particular school. We also found a substantial amount of between school variation for many sub-domain scales. The existence of such variation opens up the possibility of statistically investigating what school factors might be associated with different patterns of school leadership. We are currently engaged in exploratory analyses that attempt to reveal commonalities in leadership practices among schools at the low and high ends of CALL sub-domain scales.

Sub-domain	Teacher Level Reliability	School level Reliability	Level 1 Variance Component	Level 2 Variance Component	Proportion of Variance Between Teachers	Proportion of Variance Between Schools
1.1	0.750	0.783	0.508	0.117	0.813	0.187
1.2	0.600	0.832	0.512	0.173	0.748	0.252
1.3	0.640	0.744	0.833	0.148	0.849	0.151
1.4	0.500	0.476	0.667	0.030	0.957	0.043
2.1	0.820	0.783	0.585	0.134	0.814	0.186
2.2	0.570	0.813	0.499	0.143	0.777	0.223
2.3	0.780	0.803	0.775	0.206	0.790	0.210
2.4	0.800	0.798	0.605	0.155	0.795	0.205
3.1	0.770	0.757	0.756	0.145	0.839	0.161

3.2	0.750	0.763	0.792	0.158	0.833	0.167
3.3	0.830	0.668	0.822	0.094	0.898	0.102
3.4	0.740	0.776	1.019	0.222	0.821	0.179
4.1	0.510	0.700	0.486	0.066	0.880	0.120
4.2	0.490	0.789	0.841	0.202	0.806	0.194
4.3	0.740	0.642	0.885	0.088	0.910	0.090
4.4	0.500	0.701	0.869	0.119	0.879	0.121
4.5	0.760	0.784	0.621	0.144	0.812	0.188
5.1	0.430	0.800	0.676	0.176	0.794	0.206
5.2	0.690	0.785	0.753	0.176	0.811	0.189
5.3	0.730	0.750	0.800	0.147	0.844	0.156

Conclusions

Based on the results of focus groups, the review of survey experts, and the pilot study, the CALL survey team has undertaken a substantial revision of the survey instrument. Results from these analyses typically resulted in four types of edits to the survey: 1) Moving a question to a more appropriate sub-domain; 2) Increasing response options; 3) Clarifying wording in the question or response options; or 4) Increasing the number of questions in sub-domain. We view our work to date as part of a larger cycle in which the survey continues to be revised, and the validity of measures produced by the survey is increasingly established.