3EA | EDUCATION IN EMERGENCIES EVIDENCE FOR ACTION

INTERNATIONAL RESCUE COMMITTEE

NYU GLOBAL TIES FOR CHILDREN

Measure Guidance

Choosing and Contextualizing Assessment Measures in Educational Contexts

Silvia Diazgranados Ferráns, EdD, Senior Researcher, IRC
Jeongmin Lee, PhD, Research Advisor, IRC

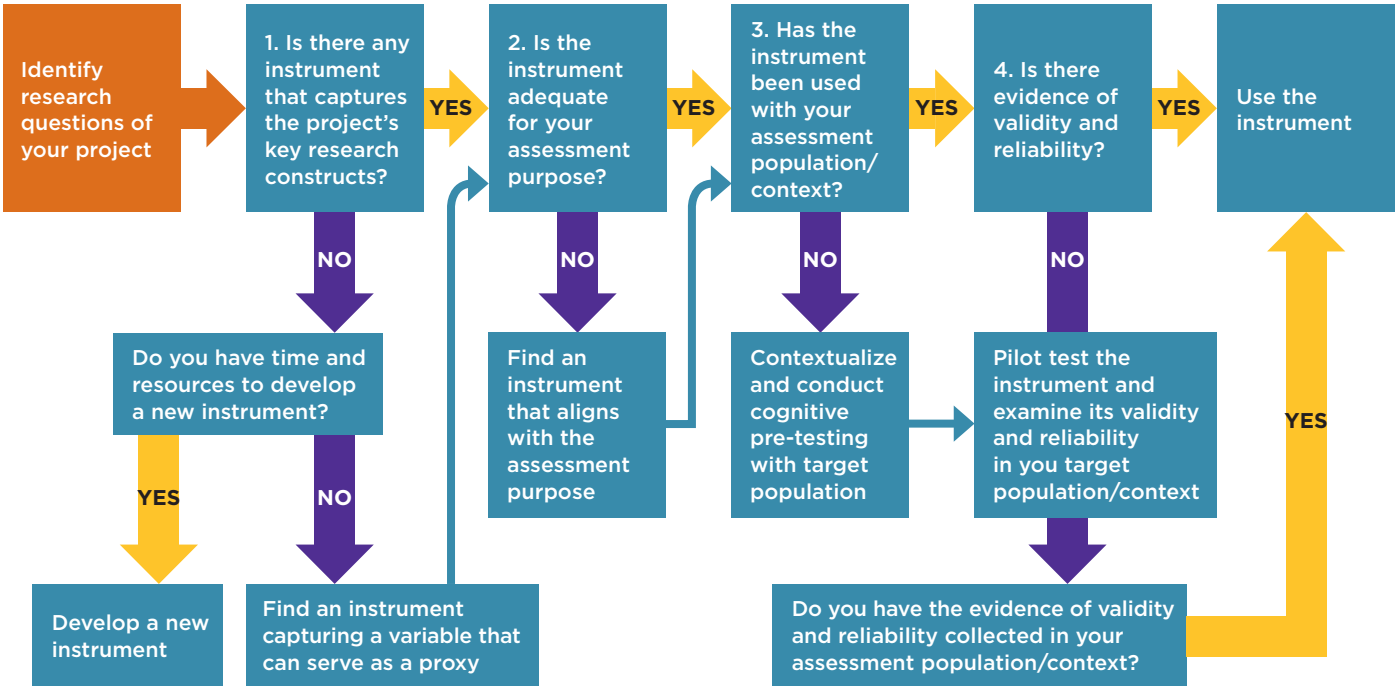
BACKGROUND

Educational assessment is the process of systematically collecting and analyzing information about characteristics of individuals, programs or settings in order to interpret the information for different purposes. These purposes may include refining programs, measuring impact and enhancing learning, achievement and healthy development. Despite growing efforts, conducting assessments in conflict and crisis setting is a remarkably challenging task given the multiplicative challenges to ensuring the accuracy, consistency and meaning of information while balancing the feasibility of assessment under extreme conditions.

In this document, we focus on the assessment of children's holistic learning and development (CHILD): specifically, the knowledge, skills and attitudes that may both promote and/or prevent children from thriving, now and in the future. We provide a step-by-step decision-making guide for researchers and practitioners interested in using CHILD measures in emergency contexts. The decision-making tree, tools and resources presented in this document will help these users make informed decisions about how to choose, contextualize and implement reliable and valid measures and understand the needs, challenges and critical support faced by the communities they serve.

PHOTO: ANDREW QUILTY/IRC

Figure 1: Decision-making tree to guide the process of choosing and contextualizing measures in contexts of conflict and crisis



STEP 1 Identify What You Want to Assess

What are you interested in assessing? Are there any particular competencies or domains you want to measure? What key construct do you want to measure and interpret? Before you can do anything else, you must be able to answer these questions and know what exactly you hope to understand. Reviewing the theory of change or log-frame of your project can help you identify the key construct that you want to assess.

Competency in education often refers to a collection of knowledge, skills, personal attitudes and behaviors which lead to the effective performance of an individual's roles or activities (Markus, Thomas, & Allpress, 2005). Examples of important competency domains for children's holistic development and learning include physical well-being, literacy, numeracy, socio-emotional processes and mental health. When a researcher interprets a learner's literacy competency based on his or her score on a test, for example, as an indicator of vocabulary knowledge, then this vocabulary knowledge is the construct that gives a meaning to observed or measured performance competency (Chapelle, 1999).

After reviewing existing instruments, you may encounter different scenarios:

- If you find an instrument that captures your construct of interest (dimension and type of competence), proceed to identify whether the instrument was designed in a way that meets your needs (See Step 2).
- If you do not find an instrument that meets your needs and you have the time and resources to develop an instrument, we suggest you do so. If you do not have the time or resources to develop that alternative, we suggest you find an instrument that captures a variable that can serve as a proxy of your construct of interest and then proceed.

Box 1. What are you measuring?

Competency types

- **Knowledge**: the ability to recall facts, information and concepts
- **Skills**: the ability to do things such as adding numbers, reading with fluency, cooperating with others to solve conflicts
- **Attitudes**: the implicit/explicit beliefs that people hold and which influence how they act
- **Behaviors**: the way people act and conduct themselves

Competency domain examples

- **Physical well-being**: competencies that may include health and hygiene, physical activity, sexual health, gross, fine and perceptual motor skills
- **Cognitive functioning**: competencies that include working memory, planning, attention control, executive functioning
- **Language and literacy**: competencies that may include letter identification, oral reading fluency, reading comprehension, writing, speaking
- **Numeracy**: competencies that may include number concepts, mathematical operations, like addition, subtraction, word problems
- **Socio-emotional processes**: competencies that may include perspective taking, emotion regulation and interpersonal processes such as conflict resolution, cooperative behavior
- **Mental health**: competencies that may include depression, anxiety, conduct problems, hyperactivity and post-traumatic stress disorder
- **Values**: competencies that may include respect of diversity, attitudes toward peace and attitudes toward gender equality

Sources: (Markus et al., 2005; Orey, 2010; Zins et al. 2004)

STEP 2 Identify an Assessment that Matches Your Needs

First, you need to consider the purpose of assessment, i.e., what are you going to use an instrument for? Various instruments are designed for different purposes: (a) describing/comparing, (b) high-stakes testing, (c) screening, (d) monitoring and (e) evaluating program impact. It is important that you identify a tool that can adequately respond to the needs of your assessment purpose because not all tools may be adequate for your purpose. For example, tools that have been developed to assess the impact of a program cannot always be used to screen individuals or track them into different trajectories. Similarly, instruments that have been developed to screen individuals are usually not adequate for the purposes of evaluating the impact of a program.

PHOTO: ANDREA FALCON/IRC

Box 2. For what purpose is the instrument designed for?

Assessment purpose

- **Identifying population needs and levels of performance**: These tools are administered to representative samples of a given population to obtain information about standard levels of performance and existing educational needs of the population. The data is often used by policy makers to address needs by directing resources accordingly.
- **Describing individual learners' performance**: Education assessments that have been designed to describe individual's levels of performance and track their progress are expected to provide meaningful feedback about learners' strengths and weaknesses in a timely way so that the learner and his/her caregivers (e.g. teachers and parents) can understand individual needs, make adjustments and provide support in a timely way.
- **Screening individuals to track them into different paths according to their needs**: Assessment tools that have been developed to screen individuals to track them into different paths require the identification of cut points—scores at/above or below which children and youth could benefit from receiving different levels of support (e.g. individuals with developmental delays or mental health difficulties, students with high levels of intelligence). The development of screening tools requires having data on large and representative samples of the population and conducting statistical analysis to identify the degree to which cut points are sensitive enough to identify all individuals who need services and/or specific enough to correctly identify individuals who do not require supports.
- **Monitoring**: Tools for the purposes of monitoring are used on a regular basis during the implementation of a given intervention to capture key activities and outputs that a program is expected to deliver. While most monitoring tools focus on directly observable indicators (e.g. attendance) others focus on more subjective dimensions (e.g. quality of instruction). While many organizations collect the data for reporting to donors and accountability, it is highly important to also use it for the purposes of conducting timely course corrections during the implementation of a program.
- **Program evaluation**: Tools that are developed to identify programming impact are used in the context of research study designs that compare the changes observed in the outcomes of a group of participants receiving an intervention with the changes in the outcomes observed among participants not receiving a treatment. For this reason, impact evaluation tools are designed to accurately capture fine-grained changes in participants' competencies and fine-grained differences in the performance of children and youth at different levels of the distribution.

Sources: (Maki, 2002; U.S. Dept of Education, 1999)

Assessment types

Additionally, we also recommend that you consider the type of assessment that best matches your needs. In education, examples of different assessment types include performance-based assessments in which examinees are asked to demonstrate their skills by answering a set of tasks, self-reported assessments in which examinees are required to provide their own assessment of their attitudes, skills or behaviors, and other-report assessments, in which a third person provides their perception of the competencies of a participant, through recollection of past interactions or with the help of observation protocols (Jupp, 2006; U.S. Dept of Education, 1999).

Each type of assessment has advantages and disadvantages. For instance, performance-based instruments can evaluate discrete and specific aspects of participants' skills in ways that are more valid than self-reports, as the latter often suffer from social desirability bias. However, using performance-based instruments can be time-consuming and often require special training and special equipment for assessors. Self-report instruments tend to be easier to administer, can capture information on attitudes that are difficult to observe directly, and are less time-consuming and less costly than performance-based assessments. However, information collected from self-reports are often biased as people tend to overestimate or underestimate their skills or provide answers that they think others will approve, especially in contexts where anonymity is not trusted (U.S. Dept of Education, 1999). Understanding the benefits and shortcomings of different instrument types, together with the specific content and competency dimension of the construct you want to measure, can help you better select, adapt or design an instrument.

PHOTO: KULSOOM RIZVI/IRC

Box 3. What type of assessments best meets your needs?

Assessment types

- **Performance-based**: Examinees are asked to demonstrate their skills by answering a set of tasks
- **Self-reports**: Examinees are required to provide their own assessment of their attitudes, skills or behaviors, usually by rating their performance across different dimensions
- **Other-reports**: A third person provides a rating of the competencies of a participant
- **Social network tools**: Participants are provided with a list of individuals within their social network, such as classroom, and asked to identify those they have a relationship with (e.g., who are your friends?) or those who meet certain attributes (e.g., who participates a lot in class?)

Sources: (Crowe et al., 2011; Maki, 2002; U.S. Dept of Education, 1999)

Based on Step 2, identify whether there is an instrument that captures your construct of interest and that has been designed to provide data for your specific purpose. After reviewing existing options, you may encounter different scenarios:

- If the instrument of your interest matches your assessment needs, you can proceed to evaluate the evidence of validity and reliability of the instrument (See Step 3).
- If none of the instruments of your interest, meet your assessment needs, we suggest that you find another instrument that captures the construct you are interested in and matches your needs. Alternatively, if you have the time and resources, we suggest that you develop a new instrument.

STEP 3 Review Evidence of Validity and Reliability for Your Target Population

Once you have identified an instrument that captures your construct of interest and adequately matches your assessment purpose and needs, you need to review whether the instrument has been previously used with your target population and/or setting (e.g., age group, grade level, gender, language, high/middle/low income, emergency/stable contexts, etc.). Identifying whether there is already evidence of validity and reliability for your population and setting of interest is critical because an instrument that accurately measures a construct in one context may function differently in another context due to socio-cultural and/or educational differences.

Box 4. How to identify whether there is evidence of reliability and validity?

Reliability is a measure of the degree to which an instrument yields the same results across multiple occasions. There are different measures of reliability, including internal reliability, inter-rater reliability and test-retest reliability.

- **Internal reliability**: The degree to which a group of items measuring the different construct deliver consistent scores
  - Check whether authors of the measure reports statistics that reflect the internal consistency of the items in the instrument, such as *Cronbach's alpha*. The values of *Cronbach's alpha* range from 0-1. Generally, values of .90 and above are considered excellent, values of .80-.90 are considered good, values of .70-.80 are considered acceptable. Values between .60 and .70 reveal weak internal consistency. Values below .60 are considered unacceptable
- **Inter-rater reliability**: The level of agreement or relative consistency between different raters or judges on the same stimulus
  - For observation tools, it is important to check the degree of agreement or consensus that exists between different raters of observers about the values they assign to the outcomes they intend to assess. *Cohen's kappa* is a useful statistic that captures inter-rater reliability. The values of *Cohen's kappa* range from 0-1. Generally, values of .81 or above are considered near perfect agreement, .61-.80 are substantial agreement and .41-.60 are moderate agreement. Values below between . 21-.40 indicate slight agreement and values below .20 indicate no agreement between raters.

Construct validity refers to the degree to which an instrument in fact measures what it claims to measure. Valid instruments provide theoretical and statistical evidence that will allow you to generalize the results from your participants' responses to their performance outside of the assessment context. An instrument can be highly reliable, and yet, not valid. Thus, ensuring that there is evidence of validity is important even when an instrument is reliable. There are many facets of an instruments' validity, and different ways of assessing them. Typically, evidence concerning the validity of interpretations of instruments address three areas:

- **Content validity** refers to the degree to which items included in an instrument are representative of and relevant to the construct to be measured. The content validity of an instrument is often achieved by consulting experts who can assess the degree to which the measure adequately captures all the dimensions of a given construct, or if there are important aspects of the construct missing.
- **Convergent and divergent validity** refer to the degree to which the items in your instrument exhibit the relationships that they are expected to exhibit. Items in the instrument should exhibit high correlations with other items that intend to measure the same construct (convergent); and lower correlations with items that intend to measure different constructs (divergent).
- **Criterion validity** refers to the degree to which your instrument correlates or is able to predict an outcome from another instrument. Criterion validity can be concurrent (i.e., outcomes collected at the same time) or predictive (i.e., outcomes collected at an earlier time are used to predict future performance —an early grade literacy assessment that predicts reading competency at later grades). In both cases, to find evidence of criterion-related validity you will need to observe a high correlation between the outcomes captured by your instrument and the outcomes captured by the external instrument.

Sources: (Huck, 2011)

After reviewing the evidence available about the reliability and validity of your instrument, you may encounter the following scenarios:

- If the instrument that captures your construct of interest has been previously used with your target population and setting and previous studies have documented with good evidence of validity and reliability, you can proceed and use the instrument. In your reports, you should always provide evidence of the reliability of the instrument in your sample, but you don't need to provide new evidence of validity.
- If there is an instrument that captures the constructs that you want to assess, but which has not yet been used in your context and/or with your target population OR which has been used with your target population but without good evidence of validity and/or reliability, we suggest you engage in a process of contextualization, piloting (See Step 4) and validation (See Step 5).

PHOTO: KULSOOM RIZVI/IRC

STEP 4 Contextualize the Instrument

If you identified the need to adjust an instrument you will need to follow a multi-step process of contextualization that can entail (a) language translation, (b) checking for semantic, idiomatic, experiential and conceptual meaning and equivalence, and (c) pilot-testing and follow-up revisions. We broadly describe such a process, and provide some exemplar methods and suggested guidelines below. However, please note the guidelines are not meant to be exhaustive.

- **Translation**: In translating an instrument to a different language, it is important to preserve conceptual equivalency across the original and translated instruments—the meaning of each item should be the same in both versions. One effective method of assuring conceptual equivalency is to translate the tool into the target language and back into the original language of the measure as this ensures appropriate and concise wording across the instruments (Chávez & Canino, 2005). To facilitate this iterative translation process, we suggest several rules to consider in the box next.

Box 5. How to translate an instrument?

- Use short, simple sentences
- Use active voice rather than passive voice
- Use terms that the typical respondent would understand
- Avoid idioms and slang
- Avoid ambiguous terms, such as probably and frequently
- Use two separate people to translate into the target language and back into the original language. Research shows that translations are more accurate when conducted by translators who are native speaker of the language into the language they are expected to translated a document.

Sources: (Beaton et al., 2000)

- **Checking for semantic, idiomatic, experiential and conceptual meaning and equivalence**: Cognitive pre-testing is one method that can help you ensure that your target population is understanding the assessment items in the way you intend. The process involves asking respondents to read the items of a test, restating in their own words the meaning of each item of an assessment, explaining in their own words what they think they are supposed to do, and sharing their answers out loud along with an explanation of why they answered in the way they did. Having access to this information can help you to identify common misunderstandings and make adjustments to more accurately capture the construct that you intend to capture. Based on the answers of participants, you can adjust the items to make them clear and comprehensible.

PHOTO: IRC

Box 6. How to conduct a cognitive pre-test?

- Find a small sample of participants (e.g., 10 people) from the target population or a population with similar attributes to the target population (e.g., similar age range, geographical location, linguistic and cultural characteristics).
- If participants can read, ask them to read the questions out loud. If the participants cannot read, ask trained enumerators to read questions to them. Then, ask the participants to explain, in their own words, how they understood the questions and whether something is unclear to them.
- If all questions are understood clearly, ask the participants to provide their answers out loud and to share their rationale for answering in a given way.
- While listening, take notes writing down questions or words that are unclear, confusing or misunderstood, so you can later identify patterns and revise items accordingly.

Sources: (Beatty & Willis, 2007; Blair et al., 2006)

- **Pilot-testing**: A pilot-test is a small-scale, trial run of an instrument with the target population under field conditions. The advantages of a pilot-test are multiple. Piloting your instrument can help you identify potential problems during the administration of the tool and take action to prevent those problems before you conduct the full-scale assessment. It can also help you have access to data to determine whether your instrument will provide you with reliable data when administered to your target population and setting. A careful selection of a pilot-test sample of participants can offer you the opportunity to gauge your target population's reaction to an instrument and to adjust the instrument before full implementation (Hassan et al., 2006). If using a tablet or cell phone for data collection, a pilot test can also identify any problems with the software that reduce user friendliness and create challenges for data cleaning and verification.

Based on the analysis of the information that you collected in the pilot study, you may encounter the following scenarios:

- If values of internal reliability and/or inter-rater reliability are low, you may want to further adjust your tool, by rephrasing items, dropping items, adding items, adjusting the scale and/or developing or improving the quality of your training materials, etc.
- If the evidence you collected during your pilot testing suggests that the tool is achieving good levels of internal consistency and/or inter-rater reliability, you can proceed to administer the tool to a larger sample and use the opportunity to conduct a validation study (See Step 5).

Box 7. How to collect and analyze pilot-test data?

Data collection

To pilot the instrument we suggest that you collect data from a minimum sample of 30-40 participants. If you are interested in comparing findings for different subgroups, we suggest you collect data from 30-40 individuals per group. Pilot-test samples need a sufficient number of participants who can adequately represent the target population. One way of achieving this condition is to include heterogeneous rather than homogenous participants.

For example, if your target group is children aged 13-18, try to include children from all ages within the range. If you expect group differences, you can stratify samples by, for instance, sex, grade or socio-economic backgrounds and then include similar numbers of participants for each stratum.

Data analysis

Analysis of pilot-test data needs to be guided by the pilot-test purpose. Regardless, we suggest that you estimate important statistics to represent the overall response pattern in the data (e.g., means and standard deviations) and measurement quality (e.g., reliability). To facilitate pilot-test data analysis, we suggest that at a minimum, you conduct the following analysis.

- **For performance-based & self-reporting questionnaires**
  - Check means, standard deviations and distribution including minimum and maximum
    - You want to have items with different difficulty levels: Easy items are those that most participants are likely to answer correctly and therefore tend to have high means. Difficult items are those that only participants with high competency levels are likely to answer correctly and therefore tend to exhibit low mean scores. Ideally, your test should have a combination of items that vary in terms of their difficulty.
  - Internal consistency reliability
    - Check *Cronbach's alpha* statistic (Recall Box 4 from above). When Cronbach's alpha is low (below .60), examine *inter-item correlations* (Do different items within an assessment/scale provide consistent scores, or do the scores vary significantly from item to item?) and *item-total correlations* (Does a given item correlate with other items or is this item uncorrelated with other items?). Generally, items that are part of the same scale exhibit strong correlations with other items and the total. A correlation value smaller than .2 indicates the corresponding item does not correlate well with the scale. We recommend that you either adjust or omit such items.
- **For observational tools conducted by different observers**
  - Inter-rater reliability
    - *Check Cohen's kappa* statistic (Recall Box 4 from above). When Cohen's Kappa statistic are low, you may want to review the items of your instrument and develop strategies to improve agreement among raters. Strategies to increase the level of agreement among different raters include adjusting or anchoring scales in objective and observable behaviors and developing high quality and standardized resources such as videos to train your observers.

Sources: (Hassan et al., 2006; Huck, 2011; McHugh, 2012)

STEP 5 Conduct a Validation Study

Once you have a tool that you believe reliably measures your construct of interest and meets your assessment purpose, you will need to collect evidence of construct validity.

After reviewing the evidence from your validation study, you may encounter different scenarios:

- If you are unable to find evidence of reliability and validity, you should make further revisions to your items.
- If you are able to find evidence of reliability and validity with your target population, you can proceed to use the instrument feeling confident about the construct validity of your data. In other words, you will able to claim that you are adequately capturing the construct that you think you are measuring. Please note that, depending on your research questions, you will still need to pay attention to other aspects of your research design not related to measurement (e.g. the internal and external validity of your research claims).

Box 8. How to find evidence of validity?

Face validity

- **Do instrument items adequately represent all important aspects of the construct?**
  - Review the key construct defined in the theory of change and log frame
  - Review how the construct is conceptualized and defined in the user's guide of the instrument or elsewhere in the literature
  - Evaluate whether the construct definition and the items of the instrument are consistent with each other
  - Have experts assess whether there are any important aspects of the construct missing or if they are well represented in the items

Convergent and discriminant validity

- **Do the items in your instrument relate to each other in the ways you expect them to be related? Do items that are part of the same subscale show stronger associations among each other (convergent validity) than among items that are part of different subscales (divergent validity)?**
  - Conduct a Confirmatory Factor Analysis (CFA), a statistical method that uses analysis of covariance to evaluate whether your data fits well with your theoretical hypothesis about the relationships between the items in your instrument. For example, the CFA will allow you to find evidence that the items that are part of the same subscale within your instrument show higher correlations among themselves than among items that are part of different subscales.

Criterion validity

- **Is your instrument able to predict outcomes based on information from external instruments?**
  - Collect data on outcomes using an external instrument for which there is already evidence of validity and reliability in your context/ population of interest.
  - Make predictions about the strength (high/low association) and direction (positive, negative) of the relationship between your instrument and the variables included in the external instrument and use the data to confirm these relationships.
  - Conduct correlational analyses between the data from your instrument and the external instrument to test your hypothesis. If you find strong associations (positive or negative) with instruments that are expected to be related to your instrument, this will serve as evidence of convergent validity. If you find no association with measures with which your instrument is not expected to have any relationship, this will serve as evidence of divergent validity.

Sources: (Holden, 2010; Huck, 2011)

CONCLUSION

Educational assessment is increasingly seen as important to improving children's holistic learning and development outcomes and refining the quality and success of programs around the world. But in the crisis and conflict settings where educational needs are often especially dire and resources very limited, educational assessments are hampered by a lack of instruments with the evidence of validity and reliability needed to conduct rigorous research studies. Given these challenges, this guide is prepared to offer researchers and practitioners in emergency settings a practical set of guidelines on how to select and contextualize instruments to measure constructs of interests in valid and reliable ways.

If you have relevant questions or comments, please contact:

Silvia Diazgranados Ferráns, Ed.D. Senior Researcher, Education, IRC at silvia.diazgranadosferrans@rescue.org.

Jeongmin Lee, PhD, Research Advisor, IRC at jeongmin.lee@rescue.org.

PHOTO: NED COLT/IRC

GLOSSARY

Cohen's kappa	a statistic that measures inter-rater reliability
Competency	a collection of knowledge, skills, attitudes and behaviors which leads to the effective performance of individual's roles or activities
Construct	a meaningful interpretation of observed performance consistency
Cronbach's alpha	a statistic that measures internal consistency reliability
Instrument	a measurement tool or device designed to obtain data on a topic of interest from research subjects, such as questionnaire, test or survey
Reliability	a measure of the degree to which an instrument yields the same results across multiple occasions
Validity	the degree to which an instrument measures what it claims to measure

FRONT COVER PHOTO: NED COLT/IRC

REFERENCES

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The Practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287–311.

Blair, J., Conrad, F., Ackermann, A. C., & Claxton, G. (2006). The effect of sample size on cognitive interview findings. *Meeting of the American Association for Public Opinion Research. Montreal, Quebec, Canada*.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Chávez, L. M., & Canino, G. (2005). Toolkit on translating and adapting instruments. Available at: Http://www. Hsri. Org/Files/Uploads/Publications/PN54_Translating_and_ Adapting. Pdf.

Crowe, L. M., Beauchamp, M. H., Catroppa, C., & Anderson, V. (2011). Social function assessment tools for children and adolescents: A systematic review from 1988 to 2010. *Clinical Psychology Review, 31*(5), 767–785.

Hassan, Z. A., Schattner, P., & Mazza, D. (2006). Doing a pilot study: Why is it essential? *Malaysian Family Physician : The Official Journal of the Academy of Family Physicians of Malaysia, 1*(2–3), 70–73.

Holden, R. R. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Huck, S. W. (2011). Reliability and Validity. In S. W. Huck (Ed.), *Reading statistics and researc*h (pp. 68–89). New York, NY: Pearson.

Jupp, V. (2006). T*he SAGE dictionary of social research methods*. London: SAGE publications.

Maki, P. L. (2002). Developing an assessment plan to learn about student learning. *The Journal of Academic Librarianship, 28*(1–2), 8–13.

Markus, L., Thomas, H. C., & Allpress, K. (2005). Confounded by competencies? An evaluation of the evolution and use of competency models. *New Zealand Journal of Psychology, 34*(2), 117.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

Orey, M. (2010). *Emerging perspectives on learning, teaching and technology*. North Charleston: CreateSpace.

U.S. Dept of Education. (1999). *Assessment, articulation, and accountability*, 1999: A Foreign language project. Retrieved from http://eric.ed.gov/?id=ED436978

Zins, J. E., Weissberg, R. P., Wang, M. C., & Walberg, H. J. (Eds.). (2004). *Building academic success on social and emotional learning: What does the research say?* New York: Teachers College Press.