Research & Methods
ISSN 1234-9224, e-ISSN 2544-0799
Vol. 32 (1, 2023)
The Ohio State University Libraries
Columbus, Ohio, USA
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw, Poland
https://ask.ifispan.edu.pl/
https://doi.org/10.18061/ask.v32i1.0002

The Many Faces of Cognitive Labs in Educational Measurement

Meirav Arieli-Attali
Psychology Department
Fordham University
Bronx, NY, NY, United States
E-mail: e-mail mattali@fordham.edu, meirav.attali@gmail.com
https://orcid.org/0009-0000-2362-6194

Irvin R. Katz
Cognitive and Technology Sciences Consulting
Hopewell, NJ 08525, United States
E-mail: irkatz@yahoo.com
https://orcid.org/0000-0002-7821-5824

Gabrielle Cayton-Hodges
Educational Testing Service
Princeton, NJ, United States
E-mail: gcayton-hodges@ets.org
https://orcid.org/0000-0002-6010-7526

Abstract Cognitive labs are becoming increasingly popular over the past decades as methods for *gathering detailed data on the processes by which test-takers understand and solve assessment items and tasks*. Yet, there's still misunderstandings and misconceptions about this method, and there is somewhat skepticism about the benefits of the method as well as lack of best practices for using it. This study's purpose was to clear out some of the misconceptions about cognitive labs, and specifically to show through theory and examples of use, the concrete benefits and best practices of cognitive labs in ***different*** stages of assessment development, ranging from early stages of conceptualizing and designing the task or item to later stages of gathering validity evidence for it. Previous literature review on the topic revealed that even the term "cognitive labs" describes different techniques,

originated in three different fields of study (Arieli-Attali, King, & Zaromb, 2011): 1) Cognitive Psychology and Artificial Intelligence research ("Think Aloud" studies, e.g., Ericsson and Simon, 1993); 2) Survey development studies ("Cognitive Interviews", e.g., Willis, 2005); and 3) software development studies ("Usability Test", e.g., Nielsen and Mack, 1994). While the latter two fields draw from the first original method, the different terminology and practices might have been the cause for skepticism and avoidance of use in educational measurement. This study maps the various ways of applying the method, shedding light on which variation can be used in which context of assessment development, in order to answer the research questions. We conclude that while it is evident that *uninterrupted think aloud* is needed for collecting response process validity, more flexible techniques may be used in contexts of usability or for assessment fairness or accessibility purposes.

Key words: Cognitive labs, thought processes, innovative assessment, validation methods

THE NEED FOR PROCESS-ELICITING TECHNIQUES IN ASSESSMENT DEVELOPMENT

There is increasing recognition within the assessment community that traditional forms of validation, which emphasize expert judgments about content ("content validity") or consistency with other measures ("convergence validity"), should be supplemented with evidence of the cognitive or substantive aspect of validity (Whitely, 1983; Linn, Baker & Dunbar, 1991; Messick, 1993, 1995; National Research Council [NRC], 2001, p. 206– 209). According to Messick, a central aspect of validity is *construct validity*, which enables the test developers to claim that the assessment is ***indeed measuring the construct it is intended to measure***. However, within construct validity, a *substantive* aspect pertains to evidence about cognitive *processes* that presumably take place in responding to test items. Messick stresses the "need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance" and suggests that one way to derive such empirical evidence is by using "think-aloud" protocols or eye movement records during task performance (1995, p. 745). This idea relates to the concept of *construct representation*, coined by Whitely (1983), which is derived from cognitive information processing paradigm, and consists of: "identifying the theoretical mechanisms that underlie task performance…….the relative dependence of task responses on the *processes, strategies*, and *knowledge stores* that are involved in performance" (1983, p. 180, emphasis added). Whitely emphasizes the need for explicit articulation of *a model of task performance* in the course of establishing construct representation, which may be included but

not required in the substantive aspect of construct validity as defined by Messick. Thus, increasingly, researchers in the field of measurement identified the need and the potential benefits in using ***process-eliciting techniques*** (i.e., techniques that elicit and track cognitive processes), since now validity also includes aspects or claims about processes vis-à-vis Messick, or explicit process models, termed *process models of task performance* vis-à-vis Whitely. To adjust to the changes in validity conceptualization, the Standards for Educational and Psychological Testing termed ***response-process validity*** to address this aspect of validity (Standards for Educational and Psychological Testing; AERA, APA, NCME, 1999).

This increased focus on the processes of performance aligns with the aspiration to link theories of cognition and learning with assessment practices (Pellegrino, Baxter & Glaser, 1999; see also, Embretson & Gorin, 2001; Gorin, 2006; Leighton, 2004; Leighton & Gierl, 2007; NRC, 2001), originally expressed by Cronbach (1957). In "Knowing What Students Know" (NRC, 2001) the National Research Council assert the need to rethink the fundamental scientific principles of current approaches to assessment, and to broaden the assessment framework to incorporate advances in cognitive sciences as well as apply the expanded capabilities in psychometrics. With that, the focus of assessment would shift and allow for measuring and interpreting more complex forms of evidence derived from student performance. In that report the authors identify limitations of the current assessments in that 1) they do not capture the kind of complex knowledge and skills that are emphasized in contemporary standards and deemed essential for success in the information-based economy of the 21st century, like organization of knowledge, problem representations, use of strategies, self-monitoring skills, and individual contributions to group problem solving; 2) current assessments have limited if any useful implication for improving learning and teaching, the ultimate goal of education reforms; 3) current assessments are "static" in that they provide "snapshots" of achievement at particular points in time, but they do not capture the progression of students' conceptual understanding over time, which is the heart of learning; and 4) current assessments yet do not fulfill to provide fairness and equity, and concerns about differences in the performance of various groups still persist (p. 26– 29).

Several different developments around the turn of the 21st century can be seen as representing such efforts to address the limitations of traditional assessments: the rise of "performance assessment" to assess wider range of skills (Council of Chief State School Officers, 1999); the turn toward classroom assessment and formative assessment to become relevant to teaching and learning (Black & Wiliam, 1998; Heritage, 2007); the development of Technology Enhanced Assessment (TEA) to allow interactivity in assessment and address the "static" concerns (Bennett, 2002); the application of Universal Design ideas to assessment to address issues

of fairness and equity for different groups (Laitusis & Cook, 2007; Thurlow et al., 2009; Thompson, Johnstone, & Thurlow, 2002); and the inclination toward diagnostic assessments that can provide specific information regarding strengths and weaknesses in student knowledge and skills. Reconceived approaches to test theory, reflecting these trends, proposed new assessment frameworks, such as: Cognitive Design System (CDS; Embretson, 1998), Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003; Mislevy, Almond, & Lukas, 2003), and Competences Assessment Programme (CAP; Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007), among others, to support these new kinds of assessments. In these frameworks, attention is drawn to other ways of validation than the traditional "post-hoc" validation, specifically including collecting validity evidence at the stage of item development to establish *process models of task performance*. In a parallel vain, Kane (1992) proposed the reconceived approach to validation that emphasized the explicit fine-grain articulation of an *interpretive argument of the scores*, to be supported by a validity argument. Contemporary approaches in psychometrics revived and developed more sophisticated diagnostic models (cf. Rupp, Templin, & Henson, 2010) to go along with these new assessment frameworks. In these models, knowledge *attributes* are identified and incorporated into the latent trait estimates, expanding beyond the unidimensional skill that is often reflected in the scores, to be replaced by multidimensional scores. All of these developments set the stage and call for the application of process-eliciting techniques such as cognitive labs.

What is a cognitive lab?

A cognitive lab is a session where a participant is invited to perform a task in a laboratory setting with an experimenter or interviewer. The participant is introduced with the task, and is expected to perform it while "thinking out-loud", that is, while saying out-loud the inner thoughts that come into their attention (without reflection or approaching the interviewer in any way), i.e., the talking is expected to be the verbalization of the "cognitive processes" that took place (thus, cognitive lab). The experimenter is observing and recording the participant's talking, but does not interfere. The underlying principle of the uninterrupted think aloud is that any verbalization produced by a subject while performing a task represents the contents of the subject's *Short-Term Memory* (Ericsson and Simon, 1993), reflecting the inner voice. It is difficult to fully understand and appreciate the method without knowing where it was originated and what it was developed for. The original method was called ***Think-Aloud method*** and it originated in experimental psychology research on thought processes since the early 1900s in Europe (Bulbrook, 1932, Watson, 1920, among others as cited in Ericsson and

Simon, 1993; Duncker, 1945), and specifically in the area of artificial intelligence after World War II in the United States (Newell & Simon, 1972). Extensive research demonstrated the validity of this approach to produce evidence on the actual cognitive processes which take place at time of performing a task (Newell & Simon, 1972; Ericsson & Simon, 1981; 1984; 1993).

Still within this approach, another complimentary feature can be added – the *retrospective probing*. Retrospective probing is an after-task session with the participant, where the interviewer asks the participant questions about the task, specifically reflecting on how they went about solving it. This component usually adds more insightful information, although the experimenter should keep in mind that the post-task interview may not necessarily reflect authentic thought processes, but rather the subject's own interpretation and reflections on their thought processes.

The method of think-aloud was adopted by other fields of study and adapted for different purposes. The adaptations were often applied with one major change – adding *intermittent questions* while the participant is performing the task. Yet, this seemingly small change is challenging the theoretical basis of the method – can one still claim that the thought processes are the authentic ones elicited by the task, or are they derailed by the intermittent questions? Since it is arguably the second option, this change can be applied only in cases where there is interest in other aspects of task performance rather than in authentic thought processes. Such interests are those which focus on detecting *problems* in the task, such as task misunderstanding or participants' explanations or suggestions to improve it. Our study revealed two main adaptations to the think aloud method: one by survey developers to pre-pilot survey questions (the cognitive interview approach; Beatty & Willis, 2007), and the other by software designers to examine software interface usability (the usability studies approach; Nielsen & Mack, 1994). Since the terminology and the actual practices are sometimes similar and sometimes different, we first suggest below a classification of the various techniques to better understand the similarity and differences between the various techniques, primarily in terms of their purposes and potential use.

Various Cognitive Lab Techniques

As explained above, cognitive lab is a term used today for various techniques developed originally in different fields of study and for different purposes. We suggest the following classification: (1) ***think aloud session*** – uninterrupted thinking aloud while performing a task; (2) ***cognitive interview*** – interrupted thinking aloud while performing a task, with intermittent probing for additional information; (3) ***retrospective probing*** – probing for additional information

after an uninterrupted or interrupted think aloud OR ***cued or stimulated think aloud*** - uninterrupted or interrupted thinking aloud while *observing* a recording – computer screen key strokes, eye tracking traces – of oneself performing a task; and (4) ***usability testing*** – thinking aloud while performing a *computer based task* to examine the usability of the interface, with intermittent probing for additional information, and with retrospective probing (an application that combines techniques 2 and 3 with adaptions).

Our mapping to these 4 types helps distinguish the different purposes and uses, as well as the origins. Technique 1, the think aloud approach is the original and most strict technique, aiming to track thought processes, and originated in experimental psychology. Techniques 1 and 2 are both ***concurrent*** techniques, i.e., are applied while the participant is performing the task, but technique 2 allows for *intermittent probing* which has a crucial effect on the purpose and outcome of applying this technique. Technique 2 is the adaptation (or relaxation) of technique 1 to contexts less interested in authentic thought processes but rather focus in task problems (often used in survey design). Technique 3 includes several different ***retrospective*** techniques and can be applied after each of techniques 1 or 2, and as such also has a different purpose to elicit participants' post-task interpretation. Lastly, technique 4 is a specific application for computer-based task which combines technique 2 and 3 with adaptions for the particular issues involved in software design. Table 1 presents the different techniques and their primary purpose and use. This classification will also help us in specifying or suggesting the appropriate technique at each phase of assessment development.

While all the techniques share out-loud verbalization of one's thoughts related to a task at hand, the only technique that actually enable tracing *authentic thought processes,* to a certain extent, is the ***uninterrupted think aloud*** method (Ericsson & Simon, 1984, 1993; see also, Ericsson & Simon, 1981, 1998; Ericsson, 2002, 2006, 2017). All the other techniques come close to thought processes, but the processes are interrupted and thus might be derailed by the probing, yet are still useful for other purposes; eliciting and obtaining knowledge representation, comprehension and misinterpretation issues, persistent misconceptions or lack of knowledge, and in particular they enable obtaining explanations and suggestions from the point of view of participants (examinees/ students) (Beatty & Willis, 2007; Chi, 1997; Nielsen & Mack, 1994). The goals of usability testing and cognitive interviews are to detect *problems* with the items or interface, less so tracing thought processes or knowledge representations per se.

Table 1: Different cognitive lab techniques

Technique	Description	Primary purpose and use
think aloud session	uninterrupted thinking aloud while performing a task	tracing authentic thought processes
cognitive interview	interrupted thinking aloud while performing a task, with intermittent probing for additional information	eliciting and obtaining knowledge representation; identifying misunderstanding, misinterpretation and misconceptions detecting fairness and accessibly issues
retrospective probing OR cued/stimulated think aloud	probing for additional information after an uninterrupted or interrupted think aloud OR uninterrupted or interrupted thinking aloud while *observing* a recording of oneself performing a task	obtaining explanations, reflections and suggestions
usability testing	thinking aloud while performing a computer-based task to examine the usability of the interface, with intermittent probing for additional information, and with retrospective probing	specific application for computer-based task – detecting problems with item/task presentation; detecting fairness and accessibly issues

As mentioned above, while the original think-aloud method was developed by experimental psychologists and artificial intelligence researchers (e.g., Duncker, 1945; Newell & Simon, 1972), the other methods are adaptations often applied by practitioners: the cognitive interview approach was primarily used by survey developers (Beatty & Willis, 2007), and the usability testing approach was developed by software designers (Nielsen & Mack, 1994). Note that the cognitive interview technique bears close resemblance to the "clinical interview" method, termed and initially developed by Piaget, and predominantly applied in mathematics education research (Ginsburg, 1997), and in wider education research (Chi, 1997). The clinical and cognitive interviews have different goals though, the former being an assessment tool by itself, and the latter being a tool to refine an assessment tool. We emphasize this distinction when we discuss the various contexts of applying cognitive labs, and when to apply which technique.

Arguments Supporting the Need for Cognitive Labs in Measurement

There are two major claims to support the need for cognitive labs in measurement. The first claim relates to ***validity*** and goes like that: since assessment scores are assumed to indicate knowledge, ability and skills, validity evidence needs to be collected directly also about the specific thought processes that are involved in responding to those assessment items. This is essentially the major claim behind

Messick's call to go beyond traditional forms of validation to include also empirical evidence about thought processes rather than just the thought outcome (the response). Embretson took this claim one step further to say that this type of validity examination can be done also at the stage of item development within the CDS framework (Embretson, 1998), later also formulated in the ECD approach (Mislevy, et al., 2003).

The second claim relates to aspects of ***reliability*** and ***construct irrelevant variance***. Reliability examination, as well as construct-irrelevant variance examination, is charged with the task to ensure that no systematic error or irrelevant variance is captured by the assessment scores. As it turns out, cognitive labs can be applied for this purpose too. This takes the form of two separate venues: one with examinees, and one with scorers. The latter was identified by educational measurement researchers who focus on human scoring (e.g., Huot, 1988, 1990), applying think aloud with human raters to better understand *rater cognition*. The former, applying cognitive labs with examinees for the sake of identifying construct irrelevant factors was a later development (e.g., Leighton & Gokiert, 2005), that came along with the wider application of cognitive labs in the item development stage. In fact, if the purpose is *only* to identify construct irrelevant factors (and not to test process models of task performance), a more relaxed version of think-aloud may suffice. For example, computer-based assessment adopted a version of *usability testing* from the field of software design, where potential users are asked to think-aloud while navigating through a computer program. This method uses think-aloud but with intermittent probing, to direct to specific features of interest. Although the intermittent probing interferes with the thought process, in the case of identifying irrelevant features of the task, it is often enough to obtain the student interpretation and comprehension of the task. We will return to this discussion later in the paper, but for the point we want to make now, it would be suffice to say that the two major needs for cognitive labs in educational measurement, collecting validity evidence and identifying construct irrelevant factors, two variations of cognitive labs are appropriate.

The major shift or transformation that occurred in the conception of validity, not only to go beyond traditional practices of content and convergent validity to practices of process validity, but also to start the validity examination in the item development stage, opened the door to a wider application of cognitive labs in measurement. Within the traditional approach to validity, validity studies were often conducted after items were piloted and full item analysis (such as item difficulty indices, item discrimination indices, reliability estimates) was obtained. When item analysis yielded results that were difficult to interpret, such as differential item functioning, researchers looked for a way to resolve it. One way was to conduct think-aloud laboratory studies, where students or potential examinees

were asked to think aloud while responding to the test items (e.g., Freedle, Kostin, & Schwartz, 1987; Gallagher & Mandinach, 1992; Katz, Bennett & Berger, 2000). The students' verbalization was recorded, and later analyzed in light of the item analysis. The motivation for applying this then-new method in validity studies came directly from the fact that educational measurement researchers were often trained psychologists, and this was one of the method applied in research on thought processes in experimental and cognitive psychology as mentioned above (Bulbrook, 1932; Duncker, 1945; Newell & Simon, 1972; Ericsson & Simon, 1981; 1984; 1993). Applying the think aloud method also in human rating studies stemmed from the same motivation, to understand the thought processes behind the judgment process that produce a score or a rating, how raters decide to give one score or another and what factors play a role in this decision, and in particular identifying the irrelevant factors. Yet, around the 1990s, application of think-aloud in validity studies were a rare sight.

With the new frameworks, such as the CDS (Embretson, 1998), and ECD (Mislevy, et al., 2003), and the reconceived approach to argument-based validity (Kane, 1992), there became a need for an explicit articulation of the validity argument, and particularly throughout the test design and development, rather than after piloting items as was conventionally done till then. At that time appeared calls to apply the method at the item development stage (Embretson & Gorin 2001; Leighton, 2004; NRC, 2001; see also Leighton, 2017; Padilla & Leighton, 2017), however, there was some reluctance to adopt the new method and even controversy about the usefulness of the method. Embretson and Gorin assert that although the "advantage of this method is in its applicability to all types of assessment tasks, including performance assessments", they indicate that the "framework has not yet been widely applied and further validation of verbal protocol measurement is needed to address the issue of reliability" (2001, p.345). The National Research Council also notes the trade-off between the richness of the verbal data and the interrater agreement in analyzing verbal protocols; the higher the former, the lower the latter. Interrater agreement is often found to be in the range of 0.6-0.7 (NRC, 2001, p.100), which may be one of the obstacles for applying the method.

Leighton (2004) addresses the issue of the scarcity of educational measurement studies that employ cognitive labs and in particular think-aloud technique, and lists three main reasons for that: 1) educational measurement specialist are not trained in cognitive psychology; 2) collecting verbal reports is costly and time-consuming, and not always practical; and most importantly, 3) verbal protocols are perceived in the assessment community as "untrustworthy", believing that students cannot be trusted to articulate what is on their minds as they solve achievement items (Leighton, 2004; p.9). In her paper Leighton addresses what she calls the "misconceptions" about the untrustworthiness of verbal reports; misconceptions

that lead to misuse and consequently to missed opportunities. Leighton (2004) stresses that knowing the *how* and *when* to use this methodology is critical to obtaining useful information, that is, an appropriate use of the method has to involve understanding of the distinction between the different techniques and the factors that affect the usefulness of verbal reports.

Whether the information obtained via cognitive labs can inform the substantive aspect of construct validity (defining cognitive models of task performance), or detecting construct irrelevant factors, there seemed to be a need expressed by several scholars in educational measurement to communicate the potential of the method and to advocate for applying it. Indeed, in the last decade there is a growing body of work that has applied the method and found it beneficial and insightful, but it is not yet known as such to the educational measurement community. Next, we review several of these studies, organized and classified by the different phases of assessment development that they were applied to. Bringing together these studies under the classification of development phase, may help researchers and practitioners learn whether cognitive lab approach is applicable in their context, and which of the techniques they should use.

USE OF COGNITIVE LABS AT DIFFERENT PHASES IN ASSESSMENT DEVELOPMENT

Within this wide range of techniques of cognitive labs as described above, there is also a wide range of uses and applications. Chronologically in the educational measurement community, cognitive labs were first applied as a *post-hoc validation examination*, and in the *context of human scoring*, to later be followed by more and more applications at the *item development stage*. In this paper, we outline the applications by the phases in assessment development in which they fit. We classify the applications of cognitive labs into four major types of contexts or phases: (1) the phase of test design and item development; (2) the phase of determining scoring; (3) the phase of post-piloting items and test; and (4) the phase of ensuring fairness and equity. These four phases seem to follow the logical steps of test development, for example, as formulated by ECD (Mislevy et al., 2003), by which domain specification and test design (Student Model or Proficiency Model) is followed by item development (Task Model) and scoring (Evidence Model), and the issue of ensuring fairness is separate. However, in practice, this is not necessarily the case, and particularly in the case of fairness and equity matters. Traditionally fairness and equity were procedures that took place after assessments were fully developed, and even piloted, due to the nature of the post-hoc validation that were customarily exercised. However, in the past few years, these procedures not only changed their nature but also were advocated to take place at the item

development phase, as part of the Universal Design framework (Johnstone, Thompson, Miller, & Thurlow, 2008). Nevertheless, we will treat the context of fairness and equity as a separate context, since it has its unique characteristics in regards to the application of cognitive labs.

Applying Cognitive Labs at the Stage of Construct Definition and Item Development

Establishing validity evidence at the item development stage is a recent endeavor (Leighton, 2017; Padilla & Leighton, 2017), which is due primarily to the reconceived approaches to validation and the new assessment frameworks, such as the Cognitive Design System (CDS; Embretson, 1998) and Evidence-Centered Design (ECD; Mislevy et al., 2003). In this context, there is a rich variety of applications, which includes different techniques to address different research questions. Under this category we identify three major applications: (a) establishing cognitive models of task performance for specific items; (b) establishing cognitive models of domain mastery for general features of items or item models; and (c) identifying construct irrelevant factors in early stages in order to refine items, primarily related to computer-based assessments. These three contexts are substantially different in their research questions and are often confused: the first requires strict uninterrupted think aloud technique, the second – semi-structured cognitive interview, and the third – usability testing that entails a combination of interrupted think aloud and retrospective probing. To establish cognitive models of task performance for specific items, the goal is to track thought processes as authentic as possible. Thus, cognitive labs in the form of uninterrupted think aloud (with or without retrospective probing) should be used, since they are able to provide the required information about thought processes and the construct to be measured as manifested in a specific task (Ericsson & Simon, 1993). The second type, cognitive models of domain mastery are in fact definition of the construct in cognitive terms (Gorin, 2007), also referred to as *proficiency model*, *construct map* (Wilson, 2005), or *student model* in the ECD framework (Mislevy et al., 2003), and these models are not task specific but rather more general, containing the components of the Knowledge, Skills and Abilities (KSA) that constitute the construct. To establish this type of models via cognitive labs, it is not necessary to use developed test items, but rather to employ several different types of items to elicit the specific components of the KSA. For this purpose, cognitive labs in the form of interviews are more suitable. The cognitive interview is a hybrid method that stems from two unrelated origins: one is the tradition of Piaget's clinical interview approach which was widely employed in mathematics education research (Ginsburg, 1997), and the other is the cognitive interviewing

technique applied in the survey development domain (Willis, 1994; 2005), which was derived from the think aloud technique, modified to identify comprehension issues. What characterizes the cognitive interview approach is a semi-structure interview, in which some of the time the student may be thinking out loud when performing a task, but most often the interviewer interrupts with probes (either pre-determined or spontaneously). Although this technique does not allow tracing authentic thought processes for a *specific task*, due to the intermittent interruptions, it does allow identifying certain aspects of knowledge representations, including student conceptions and misconceptions regarding a specific topic. In other words, this latter context of use does not intend to say something about a specific task, but rather about certain attributes or features of tasks, that constitute mastery (proficiency) in a domain (hence, cognitive models of domain mastery) (see also Chi, 1997). The third subtype deals with identifying irrelevant features in the process of item refinement. This is specifically important in computer-based tests, where assessment developers wish to ensure that the delivery format (the computer) does not introduce difficulty sources that are construct irrelevant. The cognitive lab technique that is suitable here is of the usability testing form. This includes asking the user to talk out load while interacting with the computer interface, but with interruptions to address specific concerns of the interface developer. Here too, the authentic thought processes may not be traced, but the relevant information about the usability of the interface to allow test taker to interact with the system in the way anticipated is obtained. We bring examples of studies for each of these three uses in the item development stage.

Establishing cognitive models of task performance for specific items.

In the 1990s there were studies that started to use strict think aloud to test hypotheses about cognitive models of task performance, or to develop them from the verbal reports (e.g., Baxter & Glaser, 1998; Hamilton, Nussbaum, & Snow, 1997; Katz, 1994). For example, Baxter and Glaser (1998) relied on the original method of Ericsson and Simon (1984; 1993) and Chi (1997) for obtaining verbal protocols for performance-based science inquiry tasks, and identifying matches and mismatches between task objectives (the cognitive activities that are likely to be elicited, as articulated by test developers) and the actual student thought processes as revealed in the think-aloud sessions. The analysis of the protocols allowed the researchers to develop and iteratively refine the cognitive models of task performance; models that when articulated justify and validate the interpretation of the scores on those tasks. In other words, since the goal of the research was to develop process models of task performance, these models were specified a priori (in task analysis process), and tested via the verbalization of examinees. Hamilton, Nussbaum and

Snow (1997) conducted a similar study, foreseeing the advantages of the method to obtain information towards identifying *categories of cognitive demands* that the tasks elicited and which is valuable to explicitly take into account and articulate at the onset of task development. For example, in examining science tasks, Hamilton and his colleagues identified the following cognitive demand categories: use of working memory, use of language and communication, metacognitive skills, application of prior knowledge and expectations, acquisition of new knowledge, and use of scientific processes.

Another way to define cognitive models of task performance is by identifying differences in the processes that a task elicited in people with different levels of skill and knowledge, i.e., novices vs. experts (cf., Chi, 1997). Katz (1994) applied the think-aloud method for that purpose on an early version of the computer-based NCARB Architecture licensing exam. Comparisons of the verbal data from novices and experts, along with analyses of the test-takers interaction log (log of the sequence of each mouse clicks and key strokes), led to identification of differentiate processes taken by people with different levels of skill and knowledge. These differences in processes led, in turn, to suggestions for revising tasks to better distinguish levels of expertise. In these three examples, the researchers exploited the task(s) to elicit and trace the thought processes, and in iterative process refined the tasks so that they will better elicit the thought processes of interest.

Establishing cognitive models of domain mastery for general features of items or item models.

Cognitive models of domain mastery are detailed articulation of the fine grain bits of knowledge, skills and ability that are required in order to be proficient in a domain. With the transition within the validity approach, the traditional "test specification" is gradually replaced by "proficiency models". These models do not only specify the fine grain bits of knowledge (the nodes), but also the relationships among them (the links), creating a map of proficiencies. Similar to test specification definition, domain analysis is first required, often done via a well-established curriculum in the content domain. But unique to the proficiency models is that they are often research based, that is, they are developed based on a comprehensive literature review, synthesizing a wide scope of findings relevant to the KSA of interest (e.g., Confrey & Maloney, 2010; Arieli-Attali, Wylie, Bauer, 2012; Arieli-Attali & Cayton-Hodges, 2012). Since these models are hypothetical, cognitive labs can be a good source to obtain empirical evidence to test these models. However, for this purpose, a different form of cognitive lab was often applied. Since in these cases items are not yet fully developed or the focus is not on a particular item but rather on particular components of KSA, the "ideas" for an item is examined

via the cognitive lab. This is done by an interactive student-researcher interview, led by interviewer probes, in the tradition of Piaget's clinical interviews. For example, Baxter and Witkovski (2002) used interviews with middle school students to identify students' strategies as a means to diagnose students' level of understanding. Strategies for solving the items were extracted from the interviews, in conjunction with the students' answer. The information from the protocols supported the development of a cognitive developmental model for proportional reasoning, as well as the conceptual basis on which the design of a cognitively diagnostic assessment in proportional reasoning will be based (Baxter and Junker, 2001; Weaver and Junker 2004). Similar studies (Confrey & Maloney, 2010; Arieli-Attali & Cayton-Hodges, 2012) applied cognitive interviews with students in the purpose of developing a conceptual model as basis for assessment design. Confrey and Maloney developed a detailed learning trajectories and learning maps of mathematical concepts with potential applicability for instruction; whereas Arieli-Attali and Cayton-Hodges refined a learning progression in mathematics for the purpose of creating assessment tasks linked to the levels in the progression. In particular, Arieli-Attali and Cayton-Hodges identified features of tasks that allow observing transition from one level of proficiency to another aligned with the conceptual learning progression. Using this form of cognitive lab, a semi-structured cognitive interview with students, allowed the researchers cited here to test or develop cognitive models of domain mastery.

Identifying construct irrelevant factors in early stages of item development and refinement.

Innovative computer-based tasks may introduce additional construct-irrelevant variance, unless properly designed (Dolan, Rose, Burling, Harms, & Way, 2007; Zapata-Rivera & Bauer, 2011). Developing a Technology Enhanced Assessment (TEA) requires examining different usability issues, such as: whether students know where to click in order to answer an item or go to the next item; whether they know how to use specific applications with drag-and-drop, graphing etc.; where in the screen their eyes are focusing the most and do they ignore parts of the screen; which properties of presentation (e.g., colors, resolution, font size, picture size, etc.) works best; whether student use the interactive options and how do they use it; does the interface pose higher demand on cognitive load; and other similar issues. Dolan, Goodman, Strain-Seymour, Adams, and Sethuraman (2011) used cognitive lab technique to examine the function of computer-based innovative items in comparison with "matched" traditional multiple-choice items that covered the same content standard. One of their findings indicates that students who were less computer-savvy tended to need more time to figure out what they need to do. In

their study, Dolan and his colleagues relied on the cognitive lab to reveal whether the cognitive schema students actually employ when responding to innovative items are similar to the expected cognitive pathways employed in MC item, the degree to which the enhanced functionality of computer-based items impacts student responses, and where are the places students struggle with the interface. They argue that "the potential for the cognitive lab protocol to generate both types of data—usability-related and construct-related—was regarded as a benefit rather than a confounding factor" (p. 12). Although the researchers carefully used the strict think aloud to trace uninterrupted thought processes, and then used the retrospective interview to detect usability issues, they mention a limitation of this design in that "attention [of students] repeatedly [was] drawn to features of the items that made them unique, thus encouraging interface exploration beyond what would be expected in a normal testing environment—as well as potentially distracting them [the students]" (p.62). This latter comment illuminates that the combination of think-aloud and retrospective probing to track *both* construct related (thought-processes) and usability issues *at the same session* is in fact problematic; the thought processes are influenced and derailed even just by the focus on usability issues, and therefore inferences about construct related issue are limited or misleading. In other words, the construct related and the usability issues are two different research questions and are best addressed separately, the former by an uninterrupted think aloud, the latter by a probing centered techniques, such as an interrupted think aloud and a retrospective probing.

Applying Cognitive Labs in Determining and Validating Scoring Procedure

A major application of cognitive labs is in the context of scoring open-ended responses, or studies that investigate raters' cognition (e.g., Crisp 2008; Cumming, 1990; Huot, 1988; Joe, Harmes & Hickerson, 2011; Milanovic, Saville, & Shuhong, 1996; Sanderson, 2001; Suto & Greatorex, 2008a; Suto & Greatorex, 2008b; Vaughan, 1992; Weigle, 1994, 1999). Raters' cognition consists of the strategies raters use in scoring responses, including possible effects on their rating behavior. Open-ended items, such as constructed response, expanded response essays, performance response or portfolio, require human scoring, which often adds source of error variance to the scores. The scoring procedures of constructed-response items have important implications for the validity of the test. Validation of the scoring procedures is considered as one aspect of the construct validity, namely - the *structural* aspect that "refers to the fidelity of the scoring structure to the structure of the construct domain at issue" (Messick, 1995). According to Kane and his colleagues (Kane, 1992; Crooks, Kane & Cohen, 1996; Kane, Crooks & Cohen, 1999), a validity argument is an explicit interpretive argument

that includes the assumptions made in "inferring from the scores of the test to the statements, predictions, decisions etc. taken" based on these scores (Kane, 1992, p. 13). In Crooks et al. paper, the authors argue that close examination of the scoring process is one of the links in the validation chain, and failure to ensure adequate validity of the scoring process can reduce the validity of both score interpretations and consequent decisions for some or all students. Specifically, they identify five threats to the validity of the scoring procedures: 1) Scoring fails to capture important qualities of task performance; 2) Undue emphasis on some criteria, forms or styles of response; 3) Lack of intra-rater or inter-rater consistency; 4) Scoring too analytic; 5) Scoring too holistic. Cognitive lab techniques can help in examining and reducing all five threats, primarily by applying think-aloud with raters while they are scoring responses, but also with students while responding to items, to identify components of responses that should be addressed by the scoring rubrics. Analysis of the verbal protocols can inform not only the scoring rubric development and refinement, but also decisions about rater training, which in itself may promote interrater agreement. We provide below some examples of studies that used think-aloud techniques to validate a scoring rubrics and/or to improve rater training.

Traditionally, development of the scoring rules relies on expert judgment; however cognitive lab techniques can contribute to that end either by examining raters' behavior or by exploring examinees' performance. The later was illustrated in a study by Baxter and Glaser (1998) where the analysis of verbal protocols revealed gaps in the scoring rubrics, that is, the quality of the thinking and reasoning that was evident during the think aloud session was not reflected in the scoring rubric criteria. As a result, the rubrics were revised accordingly. Another example comes from the research agenda for the TOEFL 2000 writing (Cumming et al., 2000). This agenda suggests to use think-aloud studies of examinees as they perform different writing tasks, in order to establish score meaning, because, as they argue, the think-alouds may prove useful in determining the relevant and irrelevant processes that examinees use when they take the test. Scoring rubrics that reflect *processes* are particularly difficult to develop and can be aided by think aloud data (Jonsson & Svingby, 2007). In their literature review of scoring studies, Jonsson and Svingby report that most studies describe an *outcome- or product-scoring* system, rather than a scoring which reflects processes. They cite only one paper that used a process-based scoring procedure (Osana & Seymour, 2004, as cited in Jonsson & Svingby, 2007). The authors argue that "just by providing a rubric there is no evidence for content representativeness, fidelity of scoring structure to the construct domain or generalizability" (p. 137), and that most studies address the content validity aspect provided by experts judgments, but do not concern themselves with providing evidence to support the construct validity through the

thinking processes that are reflected or not reflected in the scoring. By collecting empirical evidence about examinees thought processes and comparing it to the scoring rubrics, the first threat to validity of the scores is addressed: ensuring that scoring does not fail to capture important qualities of the task performance.

The second threat to the validity of the scores, which cognitive labs with raters can address, is the effect of *construct irrelevant factors on raters* while they are scoring a response, namely the "threat of undue emphasis on some criteria, forms or styles of response" in the scoring process (Crooks et al., 1996). Factors that can affect raters and bias the scoring, such as *halo effect* (tending to use performance on *one* aspect of the task to create and justify the final judgment), or *context effect* (comparing a previous response in the scoring of a current response) were evident in verbal protocols of think aloud studies with raters (Johnson, 2008; Orr, 2002; Vaughan, 1992). For example, Johnson studied factors that were influential in the process of scoring a portfolio according to a pre-specified rubric, and reported from the verbal protocol analysis that raters tended to use performance on *one* of the portfolio tasks to justify the final judgment for the whole portfolio, and that certain features of the criteria in the scoring rubrics dominated raters' overall judgments (p.18). Another issue that was evident in the verbal data was the different interpretations of the terms used in the rubric. Johnson found that raters resolved ambiguity in different ways, affecting the consistency of the scoring. Johnson study analyzed the verbal data in conjunction with other socio-contextual factors of the raters. A study that took similar approach is Joe, Harmes and Hickerson's (2011) examining verbal communication rating. Joe et al. defined general features of interest, (such as rater's memory, socio-cultural) and by applying think aloud techniques, the authors could identify which of the features of interest played a role in the scoring process. A relevant finding that replicated in both the above studies is that more experienced raters exhibited a different behavior than less experienced raters. In Johnson study, more experienced raters exhibit more holistic judgment and found it difficult to break down their judgment-making processes. This finding provides confirmation for the initial observation by Ericsson and Simon (1993) and others in the expert-novice domain (e.g., Chi, Feltovich, & Glaser, 1981) arguing that expert's condensed knowledge is more difficult to be articulated in a think-aloud session. In Joe et al.'s study more experienced raters were found to follow their own pre-existing cognitive framework rather than follow the rubrics. The interpretation of this finding by the authors suggests that the rubric required simultaneous consideration of multiple components, which was not easy to follow. Similar findings were obtained by Orr (2002), showing that raters tended to form global or holistic judgments of the performance even though the scoring framework was analytic. In sum, think aloud study with raters can examine the extent to which rubrics are followed by the raters, and identify occasions that are not satisfactory.

Specifically, by analyzing verbal data, too analytic or too holistic scoring rubrics (threats 4 & 5 in Crooks et al., 1996) can be detected and fixed, either by refining the rubrics or improving the training.

Rater training plays a central and crucial role in the quality of the rating and the resulting rater agreement. Cognitive labs in the form of think aloud were found useful also in informing and improving raters training. Several studies examined rater cognition in relation to rater reliability and implications for rater training (e.g., Crisp, 2008; Johnson, 2008; Milanovic, Saville, & Shuhong, 1996; Suto and Greatorex, 2008b; Vaughan, 1992). In particular, Johnson discusses the implication of detecting irrelevant factors in the scoring process to inform the training such that "discussion about the appropriate way *to balance such features* could form an important part of the initial training" (p.18, emphasis added). Suto and Greatorex (2008b) identified via think-aloud techniques strategies of raters' behavior (such as, matching between features in the rubrics and features in the response, scanning, evaluating, scrutinizing), which helped them to design future training. Similar study was conducted by Vaughan (1992), identifying different rater reading styles of essays. Crisp (2008) used strict think-alouds to examine rater strategies in scoring responses of different length (short, medium or long writing answers to geography questions). In particular Crisp used the behavior pattern reveled in the verbal protocol to better understand specific low inter-raters agreement. For example, she notes that two raters who were in low agreement with the other raters also exhibited verbally different distribution of the evaluation strategies and too much deliberation over scoring decisions relative to the other raters.

Understanding rater cognition goes beyond the validity of the scores of a particular assessment. When holistic scoring is used, for example for essay writing, often it is not entirely explicit what the exact nature of the construct is even when there is a high interrater agreement. The premise is that skilled human raters know how to identify "quality performance" (e.g., "good writing"), but the question remains as to "how do they identify it?" (or "what is encompassed in a good essay?"). Understanding the process of the rating of an essay can inform not only the training of new raters and the improvement of the rubric itself, but also the feedback to the writer and the score-reporting to the teacher in formative assessment framework (Popham, 1997). Moreover, with the move to automated essay scoring, unpacking rater considerations can inform the development of scoring models for machine scoring (e.g., Bejar, Williamson, & Mislevy 2006). Think aloud studies are a key component into rater cognition that is applied since the late 1980s – early 1990s (e.g., Huot, 1988; Vaughan, 1992; Weigle 1994) and till today.

Applying Cognitive Labs Post Item Piloting

Cognitive labs can provide supporting or refuting evidence regarding item validity after item analysis is completed. Although there are psychometric methods to identify differential item functioning (DIF), it is not always apparent from investigating the content of the items found with DIF the source or the reason for the DIF. In the early 1990s, researchers in educational measurement started to apply the think-aloud method in studies investigating DIF originated from different sources: gender differences (in math tests: Gallagher & Mandinach, 1992; Gallagher & De Lisi, 1994; in science test; Hamilton, 1999); Black-White differences in verbal analogies (Freedle, Kostin, &Schwartz, 1987); and later on also native vs. non-native speakers of the test language (Ercikan, et al., 2010; Ercikan et al., 2004), and students with disabilities (Johnstone, Bottsford-Miller & Thompson, 2006). For example, Ercikan and her colleagues used think-alouds to investigate to what extent surface characteristics of items that are identified by expert reviews as sources of DIF are supported by empirical evidence from examinee thinking processes in the English and French versions of a Canadian national assessment. In this research, the think-aloud protocols confirmed sources of DIF identified by expert reviews for only part of the items (10 out of 20). As the authors argue, the moderate agreement between think-aloud protocols and expert reviews indicates that evidence from expert reviews cannot be considered sufficient in deciding whether DIF items are biased and such judgments need to include evidence from examinee thinking processes. Results of such studies can have useful implication also in determining accommodations needed for foreign language examinees (we elaborate on this topic in the following section).

Cognitive labs have been also used to verify or critic "alignment" between test specifications and test functioning post-test administration (e.g., Ferrara et al., 2003; 2004; Leighton & Gokiert, 2005), or to examine unexplainable results in item analysis (e.g., Katz, Bennett & Berger, 2000). Methods of test development often include test specifications, articulating the construct to be measured in terms of the content domain. However, item format (e.g., multiple choice vs. open ended) or items' features (e.g., specific wording, figures) may introduce construct irrelevant variance, and threaten the validity of the tests. Several studies applied the cognitive lab approach to examine sources of construct irrelevant factors, post item piloting, illustrating the *actual responses* for item features and formats compared to the anticipated responses. Since this use involves the characterization of cognitive processes, the most suitable technique among the various options is the uninterrupted think-aloud verbal protocols, often accompanied with retrospective probing for complementary information. These applications were more commonly found for large-scale assessment program. For example, Ferrara and his colleagues

(2003; 2004) investigated the "alignment" between intended and actual cognitive demand of items for a state-wide standardized science assessment, reporting detected misalignment (in 2 items out of 20), and partially misalignment (in 11 out of 20); identifying sources of discrepancies in seemingly simple words and phrase choices that, as the author concluded, not only disrupt how students understood the nature of the task, but also derail students' cognitive processing, not anticipated by the item developers. Leighton and Gokiert (2005) conducted a similar study and reported misalignment between student interpretation and item developer intention in released items from the School Achievement Indicators Program (SAIP) Science Assessment (a national standardized large-scale science assessment administered in Canada). Katz, Bennett & Berger (2000) used the think-aloud methodology to examine validity of disclosed forms of the SAT-Math section, pertaining to different item format: constructed response (CR) vs. multiple choice (MC) test items. The motivation for the study arose from discrepancies in difficulty levels that were found for parallel items of different format. Katz and his colleagues hypothesized that the different item formats may have elicited different strategies, and for that reason they utilized the think aloud technique to characterize the strategy used for solving parallel items of different format. The data derived from analysis of the protocols refuted the hypothesis but provided insights that "instead of solution strategies mediating the effects of format on difficulty, the results suggest that *comprehension* factors mediate the effects of format on both strategy choice and difficulty" (p. 53, emphasis is added). That is, here too, analysis of the verbal data enable to identify construct irrelevant factors that pose a threat to the validity of the items, otherwise meant to be parallel. In sum, applying cognitive labs, often in the form of uninterrupted think aloud, can serve as a complementary method for psychometric item analysis, to test specific hypotheses about DIF or unexplainable results, or to verify or critic alignment.

Applying Cognitive Labs to Address Fairness and Equity Issues

Addressing the concern regarding the fairness and equity of tests among various groups, researchers and test developers over the last decade have become increasingly engaged in efforts to improve the accessibility of assessment to all students (Laitusis & Cook, 2007) and incorporating ideas of Universal Design (UD) to maximize accessibility as built-in in the assessment development (Thompson, Johnstone, & Thurlow, 2002). In ensuring UD, assessment developers need to verify that items function the same way for different groups of students, that is, that the items measure the same construct for students from different groups, and most importantly, that UD considerations need to be included at the initial process of developing an assessment, and not as post hoc accommodations. UD refers to

the concern in assessment development to maximize accessibility, fairness and validity of the assessment for *all students* (hence *Universal*), pertaining to students from various groups such as English language learners (ELL) and students with disabilities (learning or physical disabilities), and which is also built-in in the assessment design (hence *Design*) as opposed to providing accommodations retroactively in administration (Dolan & Hall, 2001; Thompson, Johnstone & Thurlow, 2002). In the domain of ensuring UD and accessibility to all students, several studies have already adopted or recommended the use of cognitive lab methods (e.g., Almond et al., 2009; 2010; Dolan et al., 2007; Haertel et al., 2012; Johnstone, Bottsford-Miller & Thompson, 2006; Johnstone, Liu, Altman, & Thurlow, 2007; Johnstone, Thompson, Miller & Thurlow, 2008; Thurlow, Laitusis, Dillon, Cook, Moen, Abedi, & O'Brien, 2009).

The research concerned with ELL was driven by the need to better understand the factors that play a role in low-performing ELL students on assessment tasks that intend to measure non-language skills, like mathematics, science etc. Studies that used think-aloud technique with ELL students (e.g. Abedi & Lord, 2001; Martiniello, 2008; Sato, Rabinowitz, Gallagher, & Huang, 2010; Shaftel, Belton-Kocher, Glassnap, & Poggio, 2006) indeed could identify specific sources of difficulties, such as struggle with specific categories of vocabulary (e.g., multiple meaning words, slang/conversation words), and also words that are often learned in an English-speaking home but are not familiar to non-English speaking home, such as weed, chores, etc. (Martiniello, 2008; Shaftel et al., 2006). Another source of difficulty that was evident in the verbal data was certain grammar difficulties, like prepositional clauses, using cognates, etc. (Martiniello, 2008). Abedi and Lord, (2001) observed that when students read test items aloud, they paused on unfamiliar words, or rephrased passive sentences to active ones. These identified strategies were obtained from the analysis of the verbal protocols and they are beneficial in suggesting ways to modify items for universal design purposes at the stage of item development. Similar findings were reported by Johnstone et al. (2006), emphasizing how the information from the verbal protocols was later useful in the refinement of the items to enhance their accessibility.

Johnstone and his colleagues (2006) used the think-aloud method with a variety of students with learning disabilities, hearing impairments, and cognitive disabilities, and also English language learners, and students without disabilities who were proficient in English. The authors concluded that the think-aloud method appeared to be effective for all populations, with the exception of students with cognitive disabilities. This finding and others (Almond et al., 2009) suggest that using the strict think aloud technique with some populations is not feasible. Almond and his colleagues provide a general recommendation towards the use of Cognitive Interview (CI), or other probing-centered technique that are more

structured and thus lead a less free-flow session to better address the needs of those students. Johnstone at al. (2006) suggest in some cases to use the Role-Play Protocol method (van Someren et al., 1994). The Role-Play method includes asking students to act like a teacher and instruct the researcher in how to solve a problem. This style is more interactive than traditional think aloud protocols, and may be a better method of gathering information from this population.

In fact, Almond et al. (2009) in a white paper regarding the applications of UD standards, proposes a multi-step design, in which CI can be used in various ways. In the first exploratory phase, even before items are developed, CI can be used to generate general hypotheses about the needs of special groups, for example regarding the effects of graphics, different formats or layouts, etc. Once items are developed, CI can be used to confirm/refute hypotheses, and to obtain information regarding the degree to which items are performing as intended. Once the test is operationally administered, CI can provide information to help the interpretation of the test results, specifically understanding item difficulty, item discrimination index and possible bias issues (for more details, see Almond et al., 2009).

The use of technology in assessment allows for inclusion of students from various groups thanks to the flexibility of interfaces and the enhanced opportunities to demonstrate relevant knowledge skills and ability (Almond et al., 2009; Thompson, Johnstone, & Thurlow, 2002). The three principles of UD that entail multiple means of *representation*, multiple means of *action* and *expression*, and multiple means of *engagement* (Rose & Meyer, 2002; as cited in Haertel et al., 2012), are more easily addressed by computer-based assessment. However, computer-based assessment can introduce unique accessibility issues including student familiarity with computers and the manipulation required to respond to items. In sum, cognitive labs, and particularly the more flexible interview approach can be used in UD studies primarily to identify construct irrelevant factors, and to refine items accordingly so that they are accessible to all in the sense that are measuring the same construct for all students from all groups.

DISCUSSION

Studies in educational measurement research already utilized cognitive lab techniques for different purposes. We identify two primary purposes, one that aims at capturing thought processes – focusing on the person's (examinee) point of view as the substantive aspect of construct validity, and the other that aims at examining item/test functioning – focusing on the assessment development point of view, in the purpose of identifying construct irrelevant factors. Although these goals are not separate, our claim is that it is important to make this distinction in order to make the best use of the specific cognitive lab technique.

Cognitive labs can be used for capturing thought processes as well as identifying knowledge representations, comprehension difficulties, or usability issues, and assessment developers are often interested in both. Capturing thought processes can be more appropriately achieved using strict concurrent think aloud with or without retrospective probing such that the elicited cognitive processes are the least interrupted. By using any of the probing-centered techniques, in which the person is interrupted with questions while working on the task, the path taken to complete the task may change as a result of the interruption, and thus information about thought processes cannot be reliably obtained. If the purpose of the assessment developers is NOT to capture the thought process per se, but rather verify that examinees understand the items as intended, or are able to navigate through the computer software presenting the items, then probing-centered techniques may yield better information than unobtrusive think aloud. Although concurrent think aloud may help identify also usability or comprehension issues, the literature indicates that it is more effective to use probing technique for that purpose.

Consequently, it is important to specify the main purpose of the study in order to choose the technique most appropriate for that purpose. Often times researchers and assessment developers are interested in both these purposes *on the same time*, that is, early in the development process of new items developers are equally interested to verify that the items indeed elicit the cognitive processes they are intended to measure, but simultaneously they wish to detect any comprehension problems or usability issues involved with the items. However, these two purposes cannot be achieved at the same time, using the same cognitive lab, and this temptation should be avoided. In using a concurrent think aloud one may track the cognitive process, but not necessarily identify usability issues, while in conducting a cognitive interview with deliberate probes to identify usability issue, the process is interrupted and so information about elicited thought processes cannot be obtained, and even if seemingly obtained may be misleading. It is important to be aware of this distinction, and to determine which of these purposes the focus of the cognitive lab is in order to obtain reliable and valid data. This distinction is profoundly illustrated by the well-known distinction in the measurement community. i.e., the one between construct validity and construct irrelevant variance.

Capturing cognitive processes elicited by items falls into the category of construct validation purposes, where the focus of the investigation is on the *person*. That is, the focus is whether the item serves as the right stimuli to elicit what it is intended to measure. This is the more basic purpose that an item developer should worry about. Intertwined with it is the question whether there are certain properties of the *item* that elicit construct-irrelevant factors, thus adding irrelevant variance to the measure. The focus in examining construct-irrelevant factors is on the functioning of the *item*. Although linked, these two questions are fundamentally different.

Logically, only after one verifies that the item possesses construct validity (i.e. measures the construct it was meant to measure), then it is appropriate to question whether additional irrelevant factors are measured unintentionally.

Specifying the focus of the study, whether it is the person thought processes, or the item functioning, is actually an explicit way to distinguish between construct validation purpose and construct irrelevance purpose, and this distinction can be of much help in choosing the right cognitive lab technique to use, think-aloud for the former whereas other adaptations for the latter. We illustrate this with an example. When translating items from one language to another, the program should be interested in verifying that construct irrelevant factor were not added in the process of translating the item. It makes sense that the translated item should not go through a construct validation process, as presumably this was already achieved at earlier stages of developing the item in the original language. The same holds for adaptation of items for specific groups, like English Language Learners or students with disabilities.

We illustrated in this paper different applications of cognitive labs in four contexts that follow somewhat a logical procedure of test development: the stage of construct definition and item development, the stage of scoring rubric development, the stage of post piloting items, and in the context of ensuring fairness and equity of tests. At each context and depending on the research questions, we discussed the appropriate technique, with examples from studies that actually applied it in this way. Although cognitive labs may have been treated with suspicious among educational measurement researchers in the past, nevertheless a large body of work accrued to prove the usefulness and benefits of these rich-data techniques.

REFERENCES

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Almond, P.J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified achievement standards (AA-MAS)*. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Arieli-Attali, M., King, T., & Zaromb, F. (2011). Cognitive Labs in the Service of Assessment and Educational Research. Unpublished literature review, Educational Testing Service, Princeton, NJ.

Arieli-Attali, M., Wylie, E. C., and Bauer, M. I. (2012). "The use of three learning progressions in supporting formative assessment in middle school mathematics" in Annual meeting of the American Educational Research Association. (Vancouver, Canada).

Arieli-Attali, M. & Cayton-Hodges, G.A. (2014). Expanding the CBAL competency model for mathematics assessments and developing a Rational Number learning progression. *ETS Research Report Series RR-14-08*; Educational Testing Service, Princeton, NJ. https://doi.org/10.1002/ets2.12008

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review*, 2(2), 114–129. https://doi.org/10.1016/j.edurev.2007.06.001

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. *Automated scoring of complex tasks in computer-based testing*, 49–82. https://doi.org/10.4324/9780415963572

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, 1(1). Available from http://www.jtla.org.

Baxter, G. P., & Glaser. R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17(3), 37–45. https://doi.org/10.1111/j.1745-3992.1998.tb00627.x

Baxter, G. P, & Junker, B.W. (2001, April). *Designing cognitive-developmental assessments: A case study in proportional reasoning*. Paper presented at the meeting of the National Council for Measurement in Education (NCME), Seattle, WA.

Baxter, G. P., & Witkowski, C. (2002). The development of proportional reasoning. Unpublished manuscript; Educational Testing Service, Princeton, NJ.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5, 7–73. https://doi.org/10.1080/0969595980050102

Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6(3), 271–315. https://doi.org/10.1207/s15327809jls0603_1

Chi, M. T. H., Feltovich, P. J. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices." *Cognitive Science,* 5(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Confrey, J., & Maloney, M. (2010). The construction, refinement, and early validation of the equipartitioning learning trajectory. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th annual conference of the learning sciences* (Vol. 1, pp. 968–975). Chicago, IL: International Society of the Learning Sciences.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, 38(2), 247–264. https://doi.org/10.1080/03057640802063486

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–684. https://doi.org/10.1037/h0043943

Crooks, T. J., Kane, M. T. & Cohen, A. S.(1996). Threats to the valid use of assessments. *Assessment in Education,* 3(3), 265–286. https://doi.org/10.1080/0969594960030302

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing." *Written Communication*, 7(4), 482–511. https://doi.org/10.1177/0741088390007004003

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL 2000 Writing Framework: A working paper. *TOEFL Monograph Series 18 (RM-00-5)*; Educational Testing Service, Princeton, NJ.

Dolan, R.P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive Lab Evaluation of Innovative Items in Mathematics and English Language Arts Assessment of Elementary, Middle, and High School Students*. Research Report. San Antonio: Pearson.

Dolan, R. P., Rose, D. H., Burling, K. S., Harms, M., & Way, W. (2007, April). *The Universal Design for computer-based testing framework: A structure for developing guidelines for constructing innovative computer-administered tests.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 300–396. https://doi.org/10.1037/1082-989X.3.3.380

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38, 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x

Ercikan, K., Law, D., Arim, R., Domene, J., Lacroix, S., & Gagnon, F. (2004, April). *Identifying sources of DIF using Think-Aloud Protocols: Comparing thought processes of examinees taking tests in English versus in French.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

Ercikan, K., Arim, R., & Law, D. Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, 29(2)*,* 24–35. https://doi.org/10.1111/j.1745-3992.2010.00173.x

Ericsson, K. A. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology*, 16(7), 981–987. https://doi.org/10.1002/acp.925

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. Hoffman. (Eds.), *Handbook of expertise and expert performance* (pp. 223–241). New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.013

Ericsson, K. A. (2017). Protocol analysis. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. pp. 425–432). New York, NY: Wiley. https://doi.org/10.1002/9781405164535.ch33

Ericsson, K. A., & Simon, H. A. (1984; 1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Ericsson, K. A., & Simon, H. H. (1981). Protocol analysis. *Psychological Review*, 87, 215–250.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Ferrara, S., Duncan, T.G., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (2004, April). Examining test score validity by examining item construct validity. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Diego, CA.

Ferrara, S., Duncan, T., Perie, M., Freed, R., McGivern, J., & Chilukuri, R. (2003, April). Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment. Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago, IL.

Freedle, R., Kostin, I., & Schwartz, L. M. (1987). A comparison of strategies used by Black and White students in solving SAT Verbal analogies using a thinking aloud method and a matched percentage-correct design. *ETS Research Report Series*, *RR-87-48*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00252.x

Gallagher, A., & Mandinach, E. (1992). Strategy use on multiple-choice and free-response items: An analysis of sex differences among high scoring examinees on the SAT-M. *ETS Research Report Series RR-92-54*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01485.x

Ginsburg, H. (1997). *Entering the child's mind: The clinical interview in psychological research and practice*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511527777

Gorin, J. S. (2006). Test design with cognition in mind. *Educational measurement: Issues and practice*, *25*(4), 21–35. https://doi.org/10.1111/j.1745-3992.2006.00076.x

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, 36(8), 456–462. https://doi.org/10.3102/0013189X07311607

Haertel, G. D., Cheng, B. H., Cameto, R., Fujii, R., Sanford, C., Rutstein, D., & Morrison, K. (2012, May). *Design and development of technology enhanced assessment tasks: Integrating evidence-centered design and universal design for learning frameworks to assess hard to measure science constructs and increase student accessibility.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, Spring*, 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hamilton, L.S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, 10, 181–200. https://doi.org/10.1207/s15324818ame1002_5

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, 89, 140–145. https://doi.org/10.1177/003172170708900210

Huot, B. (1988). *The validity of holistic scoring: A comparison of the talk-aloud protocols of novice and expert holistic raters.* Unpublished PhD dissertation, Indianna University of Pennsylvania.

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, 60 (2), pp. 237–263. https://doi.org/10.3102/00346543060002237

Joe , J.N., Harmes, J. C. & Hickerson, C. A. (2011): Using verbal reports to explore rater perceptual processes in scoring: a mixed methods application to oral communication assessment, *Assessment in Education: Principles, Policy & Practice*, 18(3), 239–258 https://doi.org/10.1080/0969594X.2011.577408

Johnson, M. (2008). Holistic judgement of a borderline vocationally-related portfolio: a study of some influencing factors. *Research Matters*: *A Cambridge Assessment Publication*, *6,* 16–18.

Johnstone, C. J., Bottsford-Miller, N., & Thompson, S. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners.* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C. J., Liu, K., Altman, J., & Thurlow, M. (2007). *Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable.* (Technical Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C.J., Thompson, S.J., Miller, N.A., & Thurlow, M.L. (2008). Universal design and multi-method approaches to item review. *Educational Measurement: Issues and Practice, 27 (1),* 25–36. https://doi.org/10.1111/j.1745-3992.2008.00112.x

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. (1992). An argument-based approach to validity. *Psychological bulletin, 112(3)*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M., Crooks, T. & Cohen. A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice* 18(2), 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Katz, I. R. (1994*). From laboratory to test booklet: Using Expert-Novice comparisons to guide design of performance assessment*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement*, 17(1), 39–57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P. (2017). *Understanding qualitative research. Using think-aloud interviews and cognitive labs in educational research*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199372904.001.0001

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26(2), 3–16. https://doi.org/10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., & Gokiert, R. J. (2005). *The cognitive effects of test item features: Informing item generation by identifying construct irrelevant variance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Montreal, Quebec, Canada.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21. https://doi.org/10.3102/0013189X020008015

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333–368. https://doi.org/10.17763/haer.78.2.70783570r1111t32

Messick, S. (1993). *Trait equivalence as construct validity of score interpretation across multiple methods of measurement*. Retrieved from: http://psycnet.apa.org/psycinfo/1993-97248-004

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Studies in Language Testing*, 3, 92–111.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behaviour of composition-markers. In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment*. Cambridge, UK: Cambridge University Press.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *Educational Testing Service Research Report Series RR 16-03;* Educational testing Service, Princeton, NJ. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

National Research Council (2001). *Knowing What Students Know: The science and design of educational measurement.* National academy press, Washington, D.C.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System 3(2),* 143–54. https://doi.org/10.1016/S0346-251X(02)00002-7

Padilla, J. L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. *Understanding and investigating response processes in validation research*. In: Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research* (Vol. 26). Cham, Switzerland: Springer International Publishing, pp., 211–228. https://doi.org/10.1007/978-3-319-56129-5_12

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the" two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, 24, 307–353. https://doi.org/10.3102/0091732X024001307

Popham, W.J. (1997). What's wrong and what's right with rubrics. *Educational Leadership, 55 (2),* 72–75.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New-York: Guilford Press.

Sanderson, P.J. (2001). *Language and differentiation in examining at A level*. Unpublished doctoral dissertation, University of Leeds, Leeds, UK.

Sato, E., Rabinowitz, S. Gallagher C., & Huang, C. (2010) *Accommodations for English language learner students: the effect of linguistic modification of math test item sets*. (Final Report. NCEE 2009-4079*)* National Center for Education Evaluation and Regional Assistance National Center for Education Evaluation and Regional Assistance.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, 11(2), 105–126. https://doi.org/10.1207/s15326977ea1102_2

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, 28(2), 9–18. https://doi.org/10.1111/j.1745-3992.2009.00143.x

Suto, W. M. I., & Greatorex, J. (2008a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, *34(2),* 213–233. https://doi.org/10.1080/01411920701492050

Suto, W. M. I., & Greatorex, J. (2008b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations." *Assessment in Education: Principles, Policy & Practice, 15(1),* 73–89. https://doi.org/10.1080/09695940701876177

Thompson, S., Johnston, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments.* Synthesis Report 44. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). Accessibility Principles for Reading Assessments. *National Accessible Reading Assessment Projects*.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes.* New York: Academic Press.

Vaughan, C. (1992). Holistic assessment: What goes on in the rater's mind? In L. Hamp Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 11–26). Norwood, NJ: Ablex.

Weaver, R., & Junker, B. W. (2004). *Model specification for cognitive assessment of proportional reasoning* (Department of Statistics Technical Report 777). Pittsburgh, PA: Carnegie Mellon University.

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197–223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological bulletin*, 93(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, Sage. https://doi.org/10.4135/9781412983655

Willis, G. (1994). *Cognitive interviewing: A "how to" guide*. North Carolina: Research Triangle Institute.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zapata-Rivera, D. & Bauer, M. (2011) Exploring the role of games in educational assessment. In Mayrath, M, Clarke-Midura, J., Robinson, D. and Shraw, G. *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Pages 147–169. Information Age Publishing.