Journal of Applied Testing Technology, Vol 23(Special Issue 1), 72-94, 2022

Using Pupillometry to Validate a KSA-Mitigated Model of Cognitive Processes

Jay Thomas

Lead Content Specialist, Science, ACT, Iowa City, Iowa, USA; jay.thomas@act.org

Abstract

A model of cognition and a construct, such as a concept map (Wilson, 2009), is critical in designing assessments of that construct. The Knowledge, Skills and Abilities (KSAs) in the construct must be put to use in order to assess what test takers know and can do (National Research Council, 2001). In order to validate a construct map for graphic literacy, a model of cognitive processes involving exerted cognitive effort and the mitigating effects of KSAs is explored. Data from pupillometry was used to quantify cognitive effort so that the KSA-mitigated model of cognition could be validated along with the construct map of cognitive processes related to graphic literacy and its assessment.

Keywords: Graphic Literacy, Index of Cognitive Activity, KSA-mitigated Effort, Pupillometry, Total Cognitive Effort

1. Introduction

One shift in the discussion of the validity of assessments over the last few years has been a greater emphasis in eliciting evidence of the cognitive processes of test takers and how those processes relate to the claims that the assessment makes about test takers. The Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], and National Council for Measurement in Education [NCME], 2014) specifically call for this evidence. In fact, they specifically mention eye movements as a relevant source of validity for some constructs (p. 15). This paper provides a method to systematically examine the cognition vertex through the use of eye tracking data, particularly the Index of Cognitive Activity (ICA) (Marshall, 2002; Marshall, 2000), which relies on pupillometry to determine the presence or absence of cognitive load that an individual is experiencing at specific points in time. This physiological measure provides information relevant to the cognition vertex that can be compared to observable behaviours (answers) and interpretations (overall score) of individuals.

1.1 Need for Evidence of Cognitive Processes for Assessment for WorkKeys Graphic Literacy

Beginning in 2015, the ACT WorkKeys Locating Information assessment underwent major revisions. The revision addressed construct validity and revised the construct to better align with the claims about what skills in graphic literacy were needed based on job profiling (ACT, 2019) and changes to the view of graphic literacy since the test was originally developed in the mid-1990s. Graphic literacy was defined as the ability to find, summarize, compare, make decisions and communicate using graphic sources such as graphs, charts, tables, process diagrams, flow charts, forms and digital dashboards. These revisions required the collection of evidence of cognitive processes to demonstrate that the proposed construct framework matched what test takers would do to complete tasks. The first validation phase began with think-alouds (Thomas

and Langenfeld, 2017a; Langenfeld, Thomas, and Gao, 2019). The think-aloud research presented evidence of the cognitive processes and the proposed framework.

The new framework greatly differed from the original construct which led to renaming the assessment to focus on the overall construct of graphic literacy rather than a specific type of one-step skill (Locating Information). However, we still needed to gather more evidence of the cognitive processes in general and the relative difficulty of the skills. To meet those goals, we turned to eye tracking (Thomas and Langenfeld, 2019; Langenfeld, Thomas, Zhu, and Morris, 2020) and pupillometry which will be the focus of this paper.

1.2 Eyetracking and Pupillometry as a Measure of Cognitive Effort

Cognition is the "black box" of the assessment triangle (National Research Council [NRC], 2001). Think-alouds give insight into what students think and do; however, this process provides significantly less information at the extremes of mental effort. For tasks that the test taker has automaticity, they offer few, if any, statements of what they are doing because the process does not require thought (Garner, 1987). For example, if you ask someone who has mastered their multiplication facts what 8 times 12 is, they will simply recall 96. However, if you ask someone who is first learning multiplication, they may think aloud that 2 times 8 is 16. Then add that 10 times 8 is 80. Finally, they add them together to get 96. On the other hand, extremely complicated tasks tend to overload the working memory of test takers and they stop using mental energy to verbalize what they are doing. (Someren, Barnard, and Sandberg, 1994; Ericsson and Simon, 1980). Additionally, thinkalouds rely on self-report of cognitive effort which can be highly subjective (Someren *et al.,* 1994) and influenced by self-efficacy (Beauchamp, 2016). During the initial think aloud research, test takers would sometimes state "that is too hard" or "I don't know nothing about science" and then state they were going to guess and move on (Thomas and Langenfeld, 2017a; Langenfeld *et al.,* 2019; ACT, 2019). Therefore, a way to objectively quantify cognitive effort was needed.

Eye tracking software provides an unobtrusive way to gather observations about cognition. Data can be acquired while an individual takes a computer-based test with no extra equipment for the test taker. An infrared sensor bar sits below the monitor recording eye movements, pupil measurements and blinks without interfering with the testing process. This is an improvement over thinkalouds as the normal test taking processes and cognition are not potentially altered by trying to articulate what the individual is doing (Someren *et al.,* 1994; Ericsson and Simon, 1980). Unlike other physiological measurements such as Functional Near-infrared Spectroscopy (FNIR) or Galvanic Skin Response (GSR) or early eye tracking studies (Hess and Polt, 1964; Beatty and Kahneman, 1966; Kahneman and Beatty, 1966), no bulky gear is required to be worn by the individual (Ayaz, Shewokis, Bunce, and Onaral, 2011; Manseta *et al.,* 2011). For paper tests, new eye tracking goggles are lightweight and allow completely free movements (Sensori Motoric Instruments [SMI], 2016; Tobii Pro, 2016).

Using eye tracking software, several important eye movements are measured, including fixations, saccades and blinks. Fixations are periods of time when the focus of the eye remains within a small area. These have been associated with periods of cognition and processing. Saccades are rapid movements from one area in the field of vision to another. They are generally associated with lower levels of cognition (Just, Carpenter, and Woolley, 1982; Andrzejewska and Stolińska, 2016). Blinks are also measured. Blinking and the rate of blinking have been studied for their relationship to cognitive effort, fatigue, and response to light (Poole and Ball, 2006; Andrzejewska and Stolińska, 2016; Tanaka and Yamaoka, 1993). Their findings have been inconclusive; as a result, blinks will not be used in this analysis.

Pupillometry has been used to measure cognitive effort for over 100 years (Schiff, 1875; Beatty and Lucero-Wagoner, 2000). The pupil dilates autonomically based on the amount of cognitive effort (Beatty, 1982; Ahern and Beatty, 1979). Early work used the diameter of the pupil as the primary measurement (Kahneman and Beatty, 1966; Beatty and Kahneman, 1966; Beatty, 1982). This was primarily due to limitations in the technology. For example, seminal work by Beatty and Kahneman relied on using cameras set to take pictures of the eye every 0.5 seconds in a darkened room. Because of the long time

between data collection, this early research relied on the diameter of the pupil as a measure of cognitive effort. This leads to several issues. First, pupil diameter does have variability among individuals (Aminihajibashi, Hagen, Foldal, Laeng, and Espeseth, 2019; Marshall, 2002). Second, the pupil constricts in response to low light (Beatty and Lucero-Wagoner, 2000). Consequently, early research that relied on taking pictures of the pupil and eye and then measuring the diameter of the pupil could be adversely affected by changes in light as the task was displayed on a screen. Later work examined the percent change in pupil diameter which attempted to correct for the variability in individual pupil size and ambient light in the room (Bailey and Iqbal, 2008).

1.2.1 The Index of Cognitive Activity (ICA)

Modern pupillometry has been greatly aided by the improvements in technology. The equipment used for this study, the SMI red-N, collects data at 60 Hz (SMI, 2016). With the greater resolution of 60 or more measurements per second depending on the hardware and software used, the measurements can move from actual pupil diameter to speed and acceleration of pupil diameter change. The dilation caused by increased cognitive effort is a small magnitude rapid reflex (Beatty, 1982; Ahern and Beatty, 1979) and the greater the speed and acceleration of that change the greater the cognitive effort (Marshal 2000, 2002, 2007). This also removes the confounding variable of individual pupil size as the rate of change is independent of the resting pupil size at the light levels of the experiment (Marshall, 2002). Moreover, the rate of change and the acceleration of that change are both more sensitive to cognitive exertion than simple percent change from previous measurement. The constriction reflex to light is a larger magnitude but slower reflex than the dilation effect (Marshall, 2000; Marshall, 2002). The Index of Cognitive Activity (ICA) software developed by Marshall uses Fourier transformations to eliminate the constriction effects and changes in lighting while computing the ICA value which measures the instantaneous cognitive load from one measurement to the next (Marshall, 2007; Marshall, Davis, and Knust, 2004; Marshall, 2000). The software calculates the percentage of time that an individual engages in cognitive activity above the threshold for each measurement recorded by the eye tracking software. Therefore, instantaneous cognitive load can be measured for thousands of intervals in a session using normal eye tracking software and equipment with sample rates of 60-500 Hz. In addition to instantaneous cognitive load during an eye movement, the average cognitive load for a task can be calculated as well. ICA does not require multiple trials like pupillary response techniques (Marshall, 2007). ICA has been shown to be reliable to justify decisions about cognitive workload and overload conditions for the military, air traffic controllers and the National Traffic Safety Bureau (Marshall, 2007; Boehm-Davis, Gray, Adelman, Marshall, and Pozos, 2003; Morrison, Marshall, Kelly, and Moore, 1997; Bartels and Marshall, 2006; Veltman and Jansen, 2006; Marshall, Pleydall-Pearce, and Dickson, 2003).

The ICA is calculated based on measurements collected by the eye tracking equipment for each measurement and calculated independently for each eye (EyeTracking, Inc, 2014). Since the dilation of each eye is connected to instantaneous cognitive load in specific parts of the brain (Marshall, 2002), it is normal for the dilation processes to be different in each eye. The average of the ICA for the left and right eye is used for the instantaneous cognitive load. The maximum value for the ICA is one and a value of zero indicates no cognitive effort beyond the calibration point which is based on focusing on dots on the screen (SMI, 2016). Additionally, since the assessment used only items with no moving graphics, the changes in lighting were minimal and accounted for by the software.

2 Models of Cognition and Effort

Based on the think aloud study (Thomas and Langenfeld, 2017a), it was apparent that at some point test takers would decide that they would not be able to solve a problem and would not exert significant mental effort. For some, it was the topic of the task (specifically, science). For others, it was unfamiliarity or complexity of the graphic. Sometimes, the totality of the task overwhelmed the test taker (Thomas and Langenfeld, 2017a; Thomas and Langenfeld, 2018). This coincides with others work on effortful or solution behaviours on assessments (Wise, 2017; Wise, Pastor, and Kong, 2009; Wise and Smith, 2011).

a. Previous Models

Young and Stanton (2001) described a model of cognition using the term mental workload of a task. It represents the total resources needed to meet performance criteria, such as completing a task and selecting the correct answer on an assessment. This mental workload is influenced by task demands, external supports and past experiences. For example, an assessment task that requires manipulation of large amounts of data may have a high task demand. However, if the individual is supplied with a task support such as a calculator, spreadsheet software or formula sheet, there would be external supports that would decrease the mental workload. They mention previous experience as a factor but do not flesh out how that specifically relates to Knowledge, Skills, and Abilities (KSAs) of individuals trying to complete a performance task.

Stephen Wise has also written about models of cognition and effort related to effortful solution behaviours (Wise and Demars, 2010). In Wise and Smith (2011), they describe resource demands of a task as a function of familiarity of the scenario, the amount of information that must be processed in working memory and the linguistic demands of the task. Later work (Wise, 2017; Wise and Kuhfeld, 2020) suggests that test takers will evaluate whether or not to exert significant effort using solution behaviours or instead use some non-solution behaviour such as guessing or skipping a task based on a self-assessment of the likelihood of success.

Other researchers (Blessing and Ross, 1996; Hmelo-Silver and Pfeffer, 2004) include task format, task complexity and time pressure as factors that impact mental load, mental effort and performance than can be predicted using task and subject characteristics.

b. A New Model: KSA-Mitigated Cognitive Effort

i. Total Cognitive Effort Required for a Task (TCERT)

The Total Cognitive Effort Required for Task (TCERT) encompasses many of the ideas mentioned previously. The cognitive complexity of the task is one component. In assessment situations, the time pressure and speedness or perceived speedness, can add to the effort required (Blessing and Ross, 1996). The amount of information that must be processed to complete the task also contributes. The complexity of the task format, such as the graphic used (Friel, Curcio, and Bright, 2001; Bryant and Somerville, 1986; Friel and Bright, 1996) or question format such as free response or multiple choice (Martinez and Katz, 1995) will contribute to the cognitive effort required. Different tasks or items will have different TCERT values. Theoretically, tasks that are deemed easier or less complex based on the construct and task specifics should have lower TCERT than more difficult constructs as shown in Figure 1. We define cognitive effort as the summation of instantaneous cognitive load (hereafter cognitive load) over time. A single event contributes cognitive effort by integrating the cognitive load over the time that it was exerted. The Total Cognitive Effort (TCE) is the summation of the cognitive efforts from the beginning of a task until its completion.

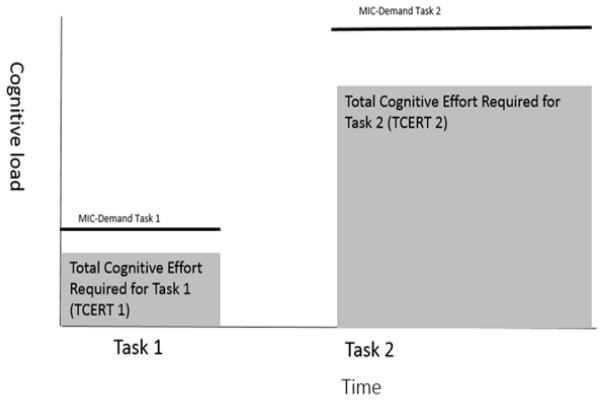

Figure 1. Model of Total Cognitive Effort for a Task (TCERT) and MIC-demand for a simple task (left) and a more complicated task right.

Consequently, TCERT could be accomplished with high instantaneous effort exerted for a short time interval or a lower instantaneous effort over a longer time interval.

Most assessment tasks require more than one step or process. Each of these subtasks will have a processing load of information that requires sufficient cognitive load to complete that task. We define here the Maximum Instantaneous Cognitive demand (MIC-demand) as the portion of a problem where the combined cognitive demands of the problem are the greatest. This is shown as a horizontal line on Figure 1 representing the maximum instantaneous cognitive demand for a given task which varies from task to task. This separation of instantaneous load and overall effort is similar to Cognitive Load Theory,

in that some of the cognitive load is directly related to characteristics of the problem while other parts of the load are related to skills that have been taught or learned (Paas, Tuovinen, Tabbers and Van Gerven, 2003; Paas and Van Merrienboer, 1994).

2.2.2 A KSA-mitigated Model of Instantaneous Cognitive Load and TCERT

Knowledge, Skills and Abilities (KSAs) represent characteristics of individuals that they use to complete tasks. Individual KSAs reduce the cognitive effort required to complete a task by using or applying previously learned knowledge and skills to the task. For example, an elementary student who is learning to multiply might exert a great deal of cognitive effort to determine the value of 12 times 12 by skip counting, using a matrix or breaking the problem into component parts. On the other hand, someone who has knowledge of either multiplication facts or perfect squares can simply recall the answer is 144 with little cognitive effort. Unique problem solving approaches have sometimes been examined as the differences between how experts and novices approach problems (Chi, Feltovich, and Glaser, 1981; Hmelo-Silver and Pfefer, 2004). KSA-mitigated Load represents the parts of a given task that can be accomplished through the application of KSAs which will reduce the mental effort required for each step of the task by using previously learned skills, processes, chunking of information and

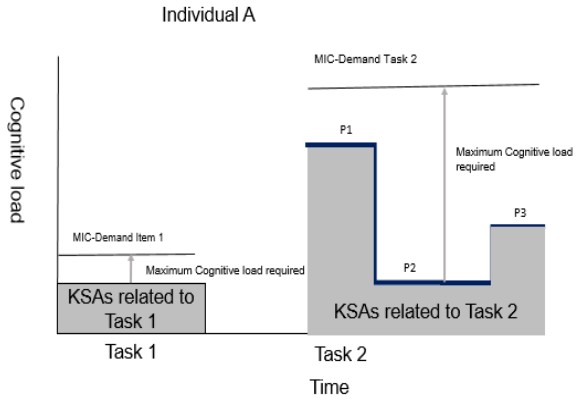

Figure 2. Model of KSAs related to a task and maximum cognitive load required.

algorithms rather than manipulation of information in working memory. The largest gap between a given set of KSAs relevant to a task and the MIC-demand would represent the Maximum cognitive load required for a given task as shown in Figure 2.

Since the task may require multiple steps, as is described later in the graphic literacy construct, the TCERT may be spread out of over multiple steps (see Task 2 in Figure 2). It is possible that an individual may have some of the KSAs necessary to complete a task, but not all of the KSAs to complete that task. A test taker may get started on a problem and get stuck. In the think aloud research (Thomas and Langenfeld, 2017a), there were several instances that test takers would get started on a problem and state that they did not know what to do next and would announce that they would need to guess and move on.

In the example above, Individual A has a set of KSAs that are relevant to each task. For Task 1, a small gap exists between the KSAs and the MIC-demand. Therefore, Task 1 has a small maximum cognitive load requirement. For Task 2, it involves three parts or steps for which Individual A has different KSAs. The maximum cognitive load required for a task would be the difference between the MIC-demand required for a task and the individual's KSAs relevant to completing the task. Additional vertical arrows could be added to represent the maximum cognitive load required for Part 1 and Part 3 of the task.

Each individual has a Personal Overload Threshold (POT), which represents the maximum instantaneous cognitive effort that an individual is able and willing to exert. POT should involve in interplay between cognitive abilities, particularly those related to working memory and metacognition and a persistence factor. This POT can be overlaid on the graph of KSA-mitigated Load (See Figure 3). Paas *et al.* (2003) referred to a construct similar to POT as "assumed cognitive capacity limit" using Cognitive Load Theory. Wise and colleagues (Wise and Smith, 2011; Wise, 2017) refer to effort capacity as the amount of effort a test taker is willing and able to give to solution (non-guessing) behaviour. As each of these increases, the POT for an individual would increase. Conversely, as these decreases, the POT would decrease as well. The sum of the KSAs and the POT represent the

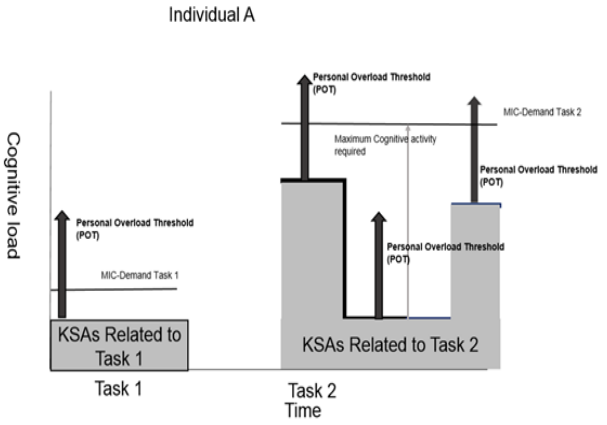

Figure 3. Model for Personal Overload Threshold (POT) for Individual A.

most cognitively complex task (highest MIC-demand) an individual can complete based on his or her particular set of KSAs relevant to the task.

If the task greatly exceeds the abilities and skills of the individual, cognitive overload occurs and performance falls off dramatically (Chen *et al.,* 2016). This may be the case for activities in a work or performance setting; in an assessment setting, if the overload threshold is exceeded, the individual will either give up or guess, because they do not have enough cognitive overload capacity and skills to complete the assessment task. Therefore, there should be a specific POT for each individual that represents the maximum instantaneous cognitive load that the

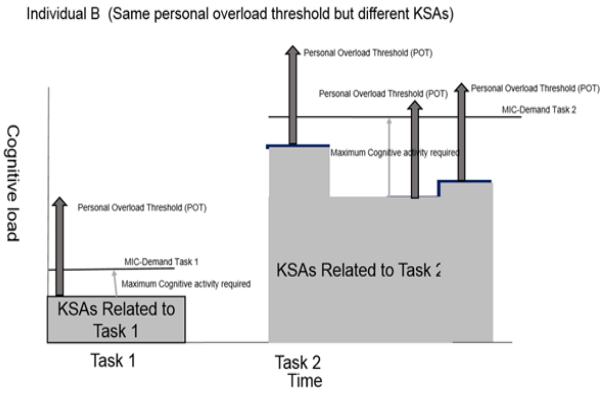

Figure 4. Model of Personal Overload Threshold (POT) for Individual B with higher KSA-mitigated load.

individual will exert before the anticipated decrease in performance occurs. Wise (2017) suggests that this results in rapid guessing behaviour which is evidence that an individual lacks the relevant KSAs for the given task even if a correct answer is selected.

Based on Figure 3, Individual A would be able to complete Task 1, because the sum of POT and KSAs related to the task exceed the MIC-demand of Task 1. Individual A would not be able to complete Task 2, because the POT is exceeded by the second step of Task 2. So, Individual A may attempt Task 2 and successfully complete the first step; however, he or she would be overloaded by the second part of the task and either guess or give up. This individual could complete the third part of Task 2, if some scaffold were given so that the overload did not occur at step two of the task. For example, if the assessment used a propagation of error scoring method like many AP exams, the student may just make up or assume an answer for part two and use that answer to complete the third part of the task that they could complete.

Consider Individual B who has greater mastery of the KSAs on the second part of Task 2 in Figure 4. Individual B would be able to complete Task 1, because the sum of the KSAs related to the task and POT exceed the MIC-demand of Task 1. Individual B would also be able to complete Task 2, because the sum of KSAs for each part of Task 2 with the POT is greater than the MIC-demand required for the three steps of Task 2. Thus, test takers with stronger KSAs in the tested construct should be able to complete more tasks with solution behaviours because the gap between MIC-demand and KSA-mitigated load will be small.

Differences in POT should also lead to differences in testing behaviour. Figure 5 illustrates Individual C with the same KSAs as Participant A but with a higher POT. This individual is able to complete both tasks. However, this individual, in completing the task, requires a greater instantaneous cognitive load than Individual B because the gap between the MIC-demand of Task 2 and the KSAs related to Task 2 is greater than that of Individual B. Conversely, if an individual had a much lower POT than Individual A, then that individual's combination of KSAs and POT would not be sufficient to solve Task 1 with its lower MIC-demand. Wise (2017) suggests that this POT will vary from task to task based on self-awareness

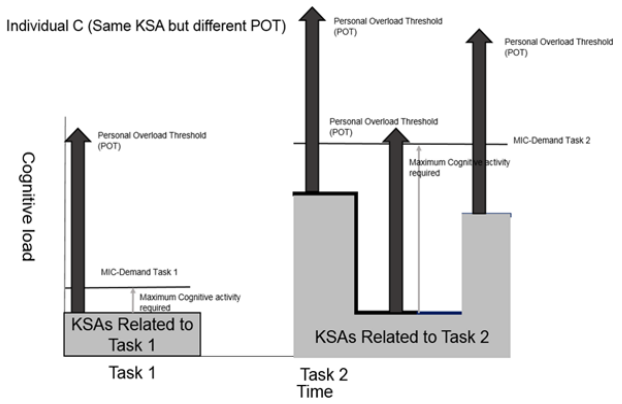

Figure 5. Model of Individual C with a High POT.

(metacognition) of a lack of KSAs referring to Kahneman's (2011) System 1 thinking as intuitively guessing rather than using effortful System 2 thinking.

Figure 5 alludes to another factor, that performance and normal workload are phenomena associated with individuals. Therefore, the mental effort for a particular task involves the intersection of the task demands and the KSAs of the individual. The task demands are related to the complexity of the information to be processed, the number and types of processes that an individual must do to complete the task, and the time allowed (Paas *et al.,* 2003). Consequently, individuals with strong KSAs for a given task should be able to reduce the total mental effort through activities like chunking of information (Chi *et al.,* 1981). Cognitive Load Theory refers to these adaptive strategies as germane or effective load (Paas *et al.,* 2003). Although Cognitive Load Theory specifically addresses issues related to teaching and instruction, the model is also relevant to assessment tasks. In a more general sense that extends beyond instruction, KSAs enable an individual to reduce the total mental effort required to complete a task. Therefore, the area represented by the KSAs on the cognitive load diagram, the KSA-mitigated effort, reduces the mental effort that must be exerted to complete the task. Using this idea, experts have KSAs that mitigate a great deal of TCERT and are able to complete tasks more efficiently and accurately than novices (Chi *et al.,* 1981). The area that is not accounted for by the KSA-mitigated effort would represent the Cognitive Effort Necessary for Task Success (CENTS). As shown in Figure 6, Task 1 has a small area of the TCERT that is not covered by the KSA-mitigated Load and therefore, a small total cognitive effort would be required to successfully attempt and complete Task 1 because the CENTS value, the area not accounted for by KSAs, is small. Conversely, Task 2 has a large CENTS area and will require a greater total cognitive effort. For any problem that an individual has some KSAs, the CENTS should be smaller than the TCERT as the KSAs will allow for some processing of information that does not require manipulation in working memory. This also allows an individual the opportunity to exert significant cognitive effort while still getting the problem incorrect.

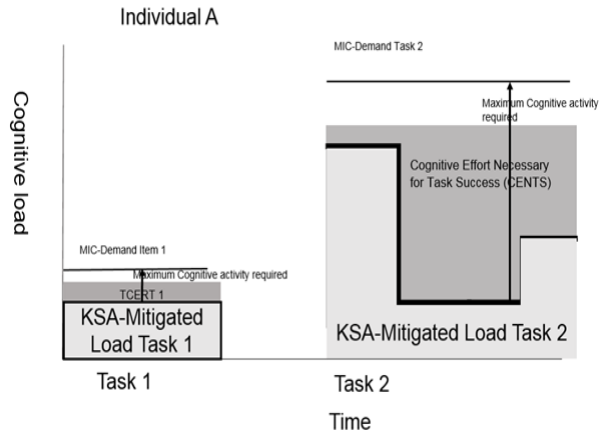

Figure 6. Model of Cognitive Effort Necessary for Task Success (CENTS) in a KSA-mitigated model.

It would be expected that given the wide range of KSAs among individuals that there would be great variability in the CENTS individuals would exert to solve tasks correctly. However, within that variability, individuals with similar overall skill levels would likely have similar CENTS for tasks. Consequently, a scatter plot of total cognitive effort for all individuals against some measure of difficulty or cognitive complexity should show a great deal of variability and a weak correlation. However, examining data within individuals there should be a natural progression of increased total cognitive effort as the cognitive complexity of a task increases. Moreover, analyzing the correlation of median or mean total cognitive effort to task difficulty or cognitive complexity should show a strong positive correlation, because the group's central tendency should represent some measure of the "average" KSAs against the cognitive demands

of the task (Marshall personal communication, Jan. 22, 2018).

c. A KSA Model and Context Model of Task Difficulty for Graphic Literacy

For the redesign of the WorkKeys Locating Information assessment, assessment designers combined multiple versions of the graphic literacy construct that defined the difficulty of a task using graphic literacy skills as a combination of the number of cognitive steps or processes that an individual must complete to successfully complete the task and the complexity of the graphic or set of graphics (ACT, 2019; Langenfeld *et al.,* 2020). Experts in graphic literacy have called simple tasks that involve locating or extracting data from graphics as reading the data (Curcio, 1987; Curcio and Artz, 1997), elementary level questions (Wainer, 1992) and extracting the data (Wainer, 1992). Tasks that require at least one step beyond locating information have been called reading between the data (Curcio, 1987; Curcio and Artz, 1997) or intermediate level questions (Wainer, 1992). Tasks that require going beyond an initial inference or understanding that requires three or more total steps such as predicting other values or using the graph to create a logical argument using claim, evidence and reasoning structure were called reading beyond the data by Curcio and colleagues and overall level of questions by Wainer. These were simply called one-step, two-step and three- (or more) step cognitive processes for the construct as defined for the assessment.

Separately, the complexity of the graphic or set of graphics contributes to the difficulty, and therefore TCERT of a task as well. Facets such as the type and familiarity of the graphic contribute to the complexity (Shah and Freedman, 2011; Shah and Hoeffner, 2002). So, bar graphs are more familiar than multiple line graphs and are less complex. Data density, complex relationships, nesting of data and number of graphics also contribute to graphic complexity (Taylor, Renshaw, and Choi, 2004; Aberg-Bengtsson and Ottosson, 2006). We found in thinkalouds that having an additional graphic added to the complexity as test takers needed to determine which graph to use and process more information (Thomas and Langenfeld, 2017a). Research showed that test takers with low graphic literacy skills would focus visual attention on the first graphic that contained words from the question stem, even if it was irrelevant. (Kliewer, Langenfeld, and Thomas, 2018; Thomas and Langenfeld, 2017a, 2017b). A four-level system of graphic complexity of Easy, Low Moderate, High Moderate, and Complex was adopted after review with external experts (ACT, 2019).

A combined model matrix was developed using both facets. The Levels in Table 1 refer to the job levels in profiling that would most likely use that skill (ACT, 2019; Lee and Nathan, 1997; Palmer and Valet, 2001). Since the assessment is linked to claims about what individuals know and can do in the workforce, the nomenclature was linked to the jobs that require those skills to be successful when first beginning a job. Others have also proposed a model that integrates facets of the graphical complexity with the task or question (Tan and Benbasat, 1990; Shah and Freedman, 2011).

Importantly, this model of the difficulty of graphic literacy tasks should be verifiable by some measure of cognitive processes, in addition to the normal evidence

Table 1. A model for the interplay and overall difficulty of graphic literacy tasks

	Graphic Complexity				
		Easy	**Low Moderate**	**High Moderate**	**Complex**
Number of cognitive steps required	One Step	Level 3	Level 3	Level 4	Level 5
	Two steps	Not Tested	Level 4	Level 5	Level 6
	Three steps	Not Tested	Level 5	Level 6	Level 7

from psychometric findings. Level 7 tasks are defined as more difficult and should require more cognitive effort than Level 6 which should require more cognitive effort than Level 5 and so on. There should be some way to actually quantify the cognitive difficulty of tasks rather than rely on psychometric properties such as the p-value or IRT-b value. Ideally, this would be quantifiable and not rely solely on self-reported measures of difficulty that can be gathered in other cognitive lab activities.

2.4 How to Measure Total Cognitive Effort (TCE)

Eye tracking software is able to calculate instantaneous cognitive load from pupillometry measurements taken during eye tracking experiments. To bridge between instantaneous cognitive load and overall mental effort the cognitive load must be integrated over the time interval over which it was calculated. The Total Cognitive Effort (TCE) an examinee exerted on an item was estimated by integrating the individual cognitive effort from the initial step of viewing the item to the final step of providing an answer. Each participant's TCE per item was estimated as shown in Equation 1.

$$\text{˘˘˘}_i = \sum_{j=1}^{n} ICA_j(\quad_j) \qquad (1)$$

where, TCE_i = the Total Cognitive Effort exerted on item i, n = the total number of eye movement measurements captured during item i, ICA_j = the ICA captured for the unique eye movement j, and $time_j$ = the elapsed time in milliseconds of eye movement j (Thomas and Langenfeld, 2018; Langenfeld *et al.,* 2020). S. Marshall confirmed the appropriateness of the TCE and she suggested that given the variability in individual's cognitive effort that examining mean or median TCEs would be acceptable for analysing item level data (personal communication, Jan. 22, 2018).

In the realm of assessment, more complex tasks, as defined by the construct, should require a greater amount of TCE. This should have two effects. First, the MIC-demand for the most complex steps of the process will increase. If the MIC-demand exceeds the POT threshold of the examinee, an overload will occur and the examinee will not be able to successfully complete the task in an intentional way. In the context of assessment, this results in either a guessing strategy or an omission, as the individual would not be able to complete the most difficult part of the task. This results in a significant drop in TCE because the examinee realizes that he or she cannot solve the task. Consequently, there will be an anticipated drop in TCE results for items that exceed the sum of POT and KSA-mitigated load. Second, for examinees with a stronger set of KSAs, the overload threshold will not be reached; however, they will likely need to exert a higher maximum cognitive load for the most difficult tasks (higher MIC-demand) as well as requiring more TCE. The TCE increase could be accomplished by maintaining instantaneous cognitive load for a longer time or increasing the instantaneous cognitive load on a unit time basis and maintaining the same effort time. It is likely that for most individuals, there will be a combination of increased maximum cognitive load (for the most difficult processes) as well as an increase in effort time (for the accumulation of multiple processes for a complex task).

3. Research Questions

- Does the integration of ICA and time yield a meaningful measure of Total Cognitive Effort (TCE)?
- Does the TCE increase for a participant as the difficulty (anticipated TCERT) of a task as defined by the construct increases?
- Does the TCE indicate that for participants who have low KSAs for a given task a tendency to cognitively disengage and guess?
- Does individual random variation in TCE render measures of central tendency uninterpretable?

4. Methods

This study utilized eye tracking methodology to analyze participants' responses to items on assessments measuring workforce-related skills. Each participant completed an assessment of workplace graphic literacy skills. All

individuals were compensated for their time after signing a participant consent form. All participants were recruited from a metropolitan area in one Midwestern state. Individuals were recruited from previous study email lists, word of mouth and through electronic backpack messages through local high schools (both traditional and alternative). Flyers were also posted at some local grocery stores.

4.1 Eye Tracking Equipment

The SensoMotoric Instruments (SMI) Redn 250 collection system set at 60 Hz collection was used for this research. The infrared bars were placed below 27-inch Asus PB278Q monitor located approximately 70 cm in front of the participant. The SMI Experiment Center 3.7 software was used and all participants successfully completed a 5 point calibration to a spatial resolution of 0.05 degrees (SMI, 2016). The supervisor monitored the system to ensure that the eye measurements were in frame. All participants completed the assessment in the allotted operational time of 55 minutes.

The eye tracking technology collected gaze data as the participants worked through the assessment. The gaze data included information on fixations, saccades, blinks and click stream data. A sampling rate of 60 Hz meant that eye movements were measured every 16.7 ms. Cognitive load was determining using EyeTracking, Inc. Workload RT V3 Academic software (Marshall, 2007). This module functions as the Index of Cognitive Activity (Marshall, 2002) and provides an ICA value for each eye separately and the bilateral ICA for each event in the data set. The ICA was calculated separately for each individual using the ICA software module. The data for each individual were exported as a comma delimited file that was then analyzed in Microsoft Office 365 ProPlus Version 15.0.4953.10001.

Eye tracking data were analyzed using SMI BeGaze 3.7 software (SMI, 2016). Since the testing platform utilizes only one URL for the entire testing sequence, a custom trial for each item for each participant was created by selecting the portion of the data that the participant worked on a given question. This was defined operationally as starting with the first screen frame in the playback that included the entire question stimulus, stem, and foils and ending after the participant used the mouse to select a response and click next. Data between questions while the software refreshed the screen for the next question were not analyzed. The load time between questions was typically between 2 and 6 seconds, although some participants experienced a lag time between some questions of approximately 10 seconds. One individual, Participant 4, was excluded from group analysis because too large a percentage of data was missing due to head movements in and out of frame so that 39% and 51% of the data were missing from the left and right eye, respectively. Data for individual items were discarded as random guessing if the individual spent less than 15 seconds to answer any item above the lowest Level 3 items.

4.2 Assessment Used

The ACT WorkKeys Graphic Literacy assessment (ACT, 2019) was designed to measure the skills individuals use when they read and comprehend graphical materials to solve work-related problems. The assessment consisted of 32 scored items covering 14 unique graphics. Embedded within the assessment form were an additional six pretest items based on three additional graphics. Graphics and items were presented together on the screen. For a few items, the size of the graphic necessitated that the examinee used the scroll feature of the computer screen to view the entire graphic along with the item. Examinees were allowed 55 minutes to complete the assessment and none of the study participants required the entire 55 minutes to finish. The assessment had been field tested and used as a part of a national workforce assessment program. The reliability of both raw scores and scale scores was strong with estimates of coefficient alpha of 0.85 or above (ACT, 2019). Participant responses were exported to a data file to assign item and overall scores for each participant. This assessment is used as part of the NCRC Work Readiness certificate program that allows test takers to document that they have skills appropriate to a variety of jobs in the Job Pro database (ACT, 2019).

4.3 Participants

The group included 10 female and 8 male participants who were recruited from a Midwest metropolitan area of approximately 170,000 people. The ethnicity of the group included one Asian, two African American and

15 white/Caucasian participants. Nine participants were enrolled in high school; three were college students; six were workforce eligible adults. Three participants were employed full-time, eight were employed part-time and seven were not employed outside the home (six high school students and one adult). Of the nine adults, the highest educational attainments were: One high school graduate, three some college with no degree, one bachelor's degree and three graduate degrees. Since the WorkKeys assessments are used in many state employment centres in addition to being given to high school students, this mix of individuals was deemed appropriate as similar to the overall testing population of the WorkKeys program.

5. Results

5.1 Assessment Scores

Individuals' answers were extracted and used to calculate the raw and scale score for the 32 operational items. Raw scores ranged from 10 to 32 items correct out of the 32 items. The mean score was 24.25 correct (SD = 7.51) and the median was 25. Four participants (22%) earned a Level Scale Score of 3 (lowest passing) or lower which is similar to national test administrations that have approximately 23% in those score ranges (ACT, 2019). These were classified as Low Scorers. Five of the 18 participants (27.8%) earned a Level Scale Score of 4 or 5 and classified as Middle Scorers. This percentage is lower than average national administration where approximately 48% of test takers earn a Level 4 or 5. The remaining 9 participants (50%) earned a Level Scale Score of 6 or 7 and classified as High Scorers. This percentage is higher than the national average of 28%.

5.2 Cognitive Effort Results for Items and Classification of Items

The TCE was calculated for each individual for each item separately, an example is shown in Figure 7 and those values were used to calculate the mean and median TCE value for each question as shown in Table 2, which includes 32 operational and 6 pretest items. Figure 7 shows the TCE increasing over time showing the cumulative cognitive effort on the task. Changes in slope are indications of changes in the instantaneous ICA readings. Steep slopes, such as the region between approximately 19,000 ms and 25,000 ms are indications

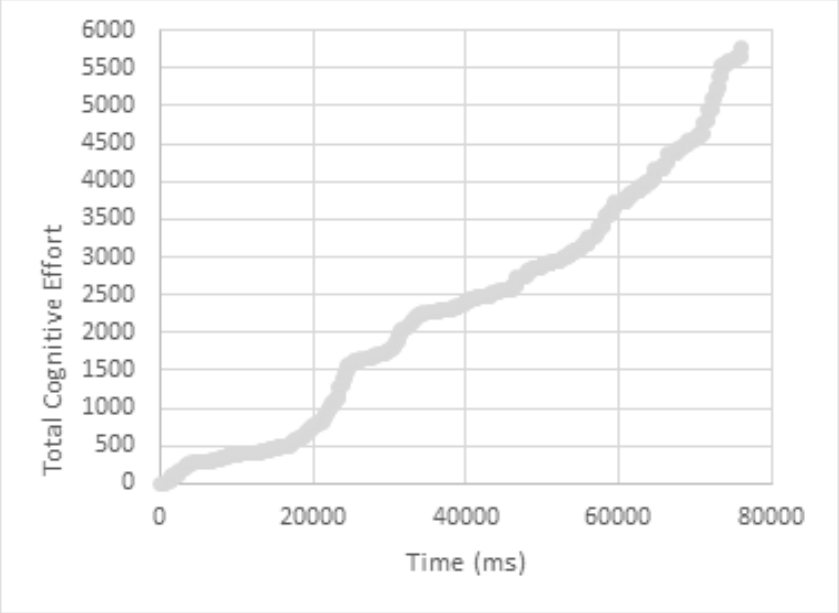

Figure 7. Example of item Total Cognitive Effort (TCE) as a function of time (Item 33, Participant 3).

Table 2. Total Cognitive Effort (TCE) descriptive statistics for each graphic literacy item

Item #	Graphic Complexity	Cognitive Steps	Overall Level	Mean	Median	SD
1	Simple	1	3	4412.1	1905.7	4031.9
2	Simple	1	3	6450.4	3097.2	6574.7
3	Simple	1	3	5041.8	2308.4	5143.7
6	Low Mod	1	3	4605.7	2735.2	4926.8
7	Low Mod	2	4	8820.8	3794.6	9599.1
8	Low Mod	2	4	9809.7	6718.3	8791.8
9	Low Mod	2	4	5416.4	2595.2	6013.5
10	Low Mod	2	4	8225.5	5078.1	7796.4
11	Low Mod	2	4	9146.1	5750.2	8300.4
12	Low Mod	2	4	7035.1	5700.4	8126.1
13	Low Mod	3	5	9997.1	8363.6	10076.8
14	High Mod	1	4	8865.2	5738.9	7441.2
15	High Mod	2	5	4671.2	3935.3	3738.4
16	High Mod	2	5	11038.8	8249.5	7604.4
17	High Mod	2	5	7217.3	3384.0	7260.8
18	High Mod	2	5	5908.1	4315.5	4521.2
19	High Mod	2	5	8778.8	3319.3	9868.2
20	High Mod	2	5	15698.2	11208.2	12249.3
21	High Mod	2	5	8094.5	5564.4	6730.6
22	High Mod	3	6	11025.2	6000.9	15726.4
23	High Mod	2	5	18311.9	13211.4	16072.4
28	Complex	2	6	7661.9	4985.8	7561.8
29	Complex	2	6	10759.3	3552.7	20264.3
30	Complex	2	6	7924.3	6500.4	6966.3
31	Complex	2	6	7308.8	6506.9	6249.7
32	Complex	3	7	12589.3	9040.7	9666.3

33	Complex	2	6	18235.5	18309.9	12706.7
34	Complex	3	7	12485.6	9363.0	10882.5
35	Complex	3	7	18090.2	14845.5	16838.0
36	Complex	2	6	7671.9	5047.2	6421.9
37	Complex	3	7	17399.0	12975.6	12609.9
38	Complex	3	7	17927.0	11552.6	16538.6

of high instantaneous cognitive load, in this case for an approximately five second period. Table 2 shows both the median and the mean. As was suggested earlier, the mean shows a tendency to skew if one or two individuals spend a great deal of time and cognitive effort.

5.3 Results based on Scoring Category (High, Middle, Low) by Portions of the Graphic Literacy Construct

Additionally, the values were clustered by graphic complexity (EASY, LOW MODERATE, HIGH

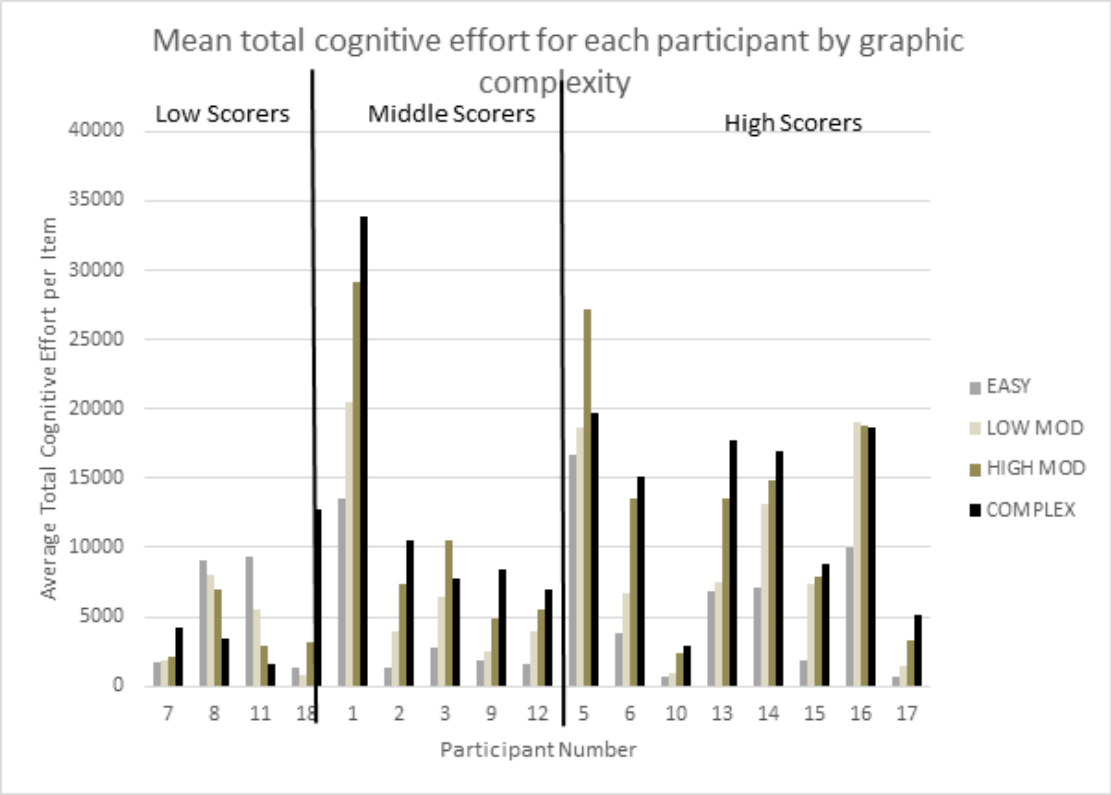

Figure 8. Mean Total Cognitive Effort (TCE) for each participant by four levels of graphic.

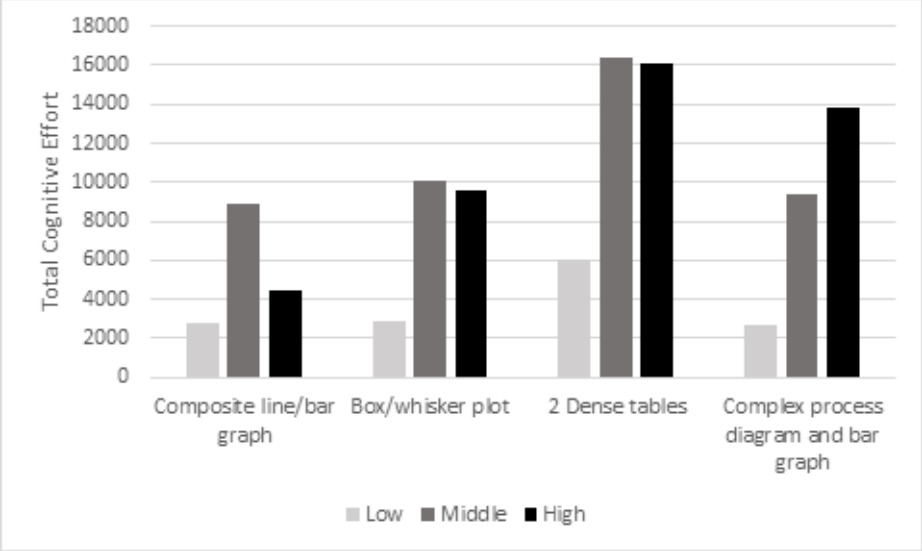

Figure 9. Mean of item median Total Cognitive Effort (TCE) for complex graphics identified by graphic type for the three different score groups.

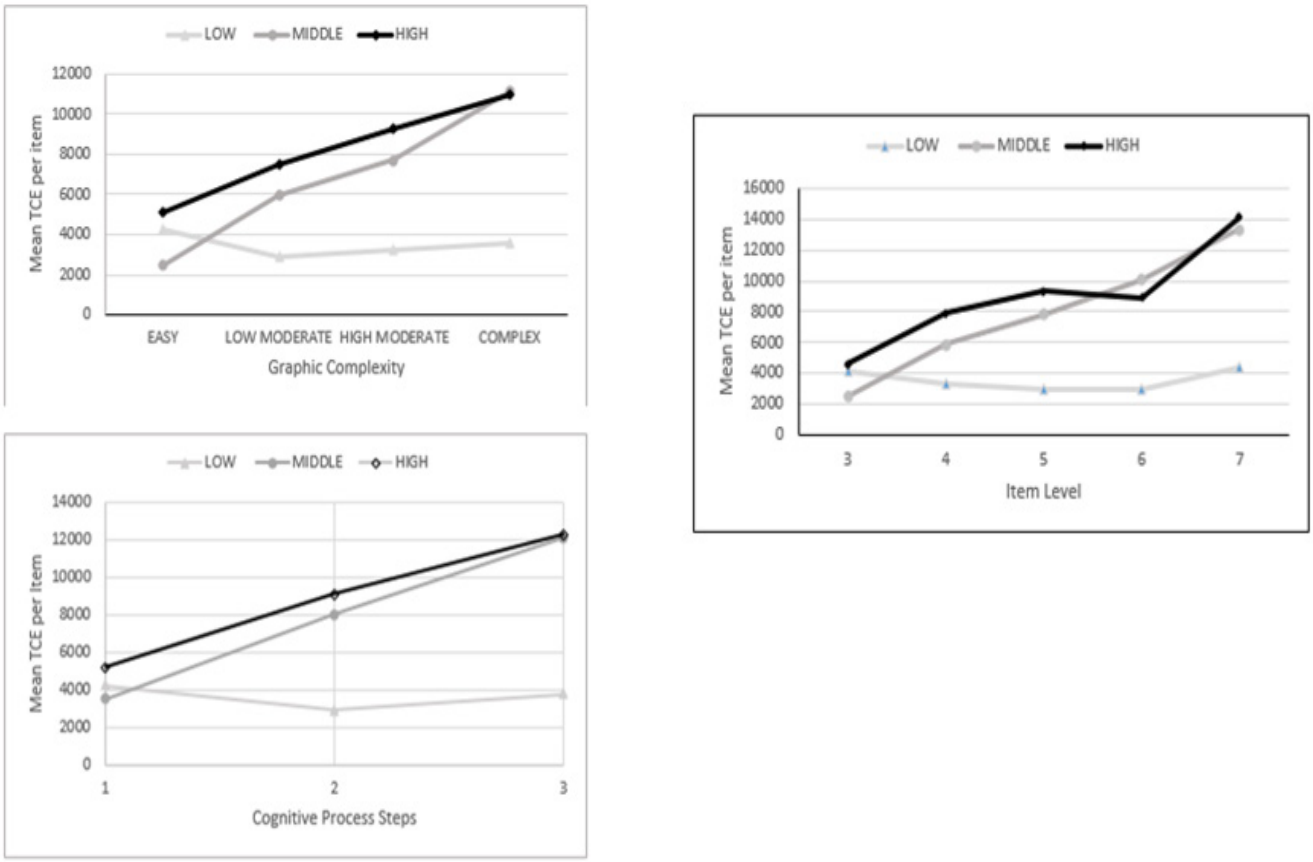

Figure 10. Mean TCE per item for the Low Scorers, Middle Scorer and High Scorers based on components of the construct: graphic complexity, number of cognitive steps, and overall item level.

MODERATE and COMPLEX) to determine relationships between graphic complexity and the total cognitive effort exerted by test takers. These results are shown in Figure 8.

For the COMPLEX graphic category, the results were also plotted by the type of graphic for a consideration of familiarity with the graphic type in Figure 9.

These results were then grouped by scoring band (High, Middle and Low Scorers) and the mean TCE value for each question that fell into that category was calculated. The top portion of Figure 10 shows the mean TCE as a function of Graphic Complexity. The middle portion shows the mean TCE as function of the number of cognitive steps required to solve the problem. Finally, the lower portion shows the mean TCE as a functional of the overall Item Level, which should represent the overall difficulty of the item.

5.4 Comparing a Low, Middle and High Scorer

The individual TCE plots for individuals can be compared when plotted on the same graph. Two example items are plotted below for the same three individuals. Figure 11 shows the TCE graphs for a Low, Middle and High Scorer for a Level 4 item. The Middle and High Scorer both answered this item correctly while the Low Scorer did not.

Figure 12 displays the TCE on a Level 6, more difficult, item that all three individuals answered correctly.

6. Discussion

6.1 Research Question 1: Does the Integration of ICA and Time Yield a Meaningful Measure of Total Cognitive Effort (TCE)

Participant response data gathered through eyetracking analysis and using the ICA software appear to support the integration. Meaningful integration was possible for all but one participant. Items that were judged to be more cognitively difficult based on the construct definition showed overall higher levels of TCE in Table 2 and the three portions of Figure 10.

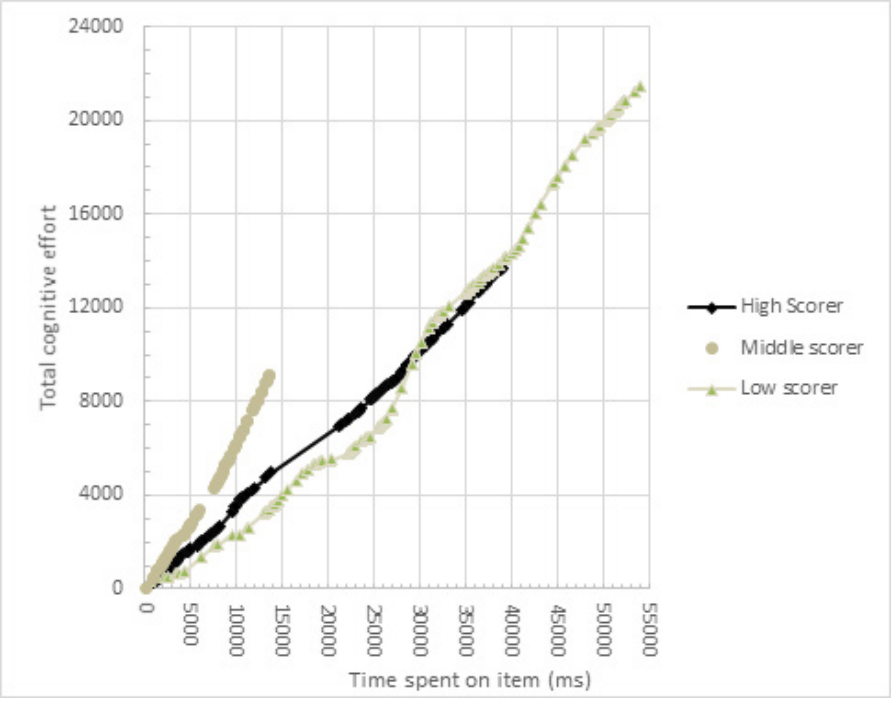

Figure 11. Comparing Total Cognitive Effort (TCE) for High, Middle and Low Scorer on a Level 4 Item.

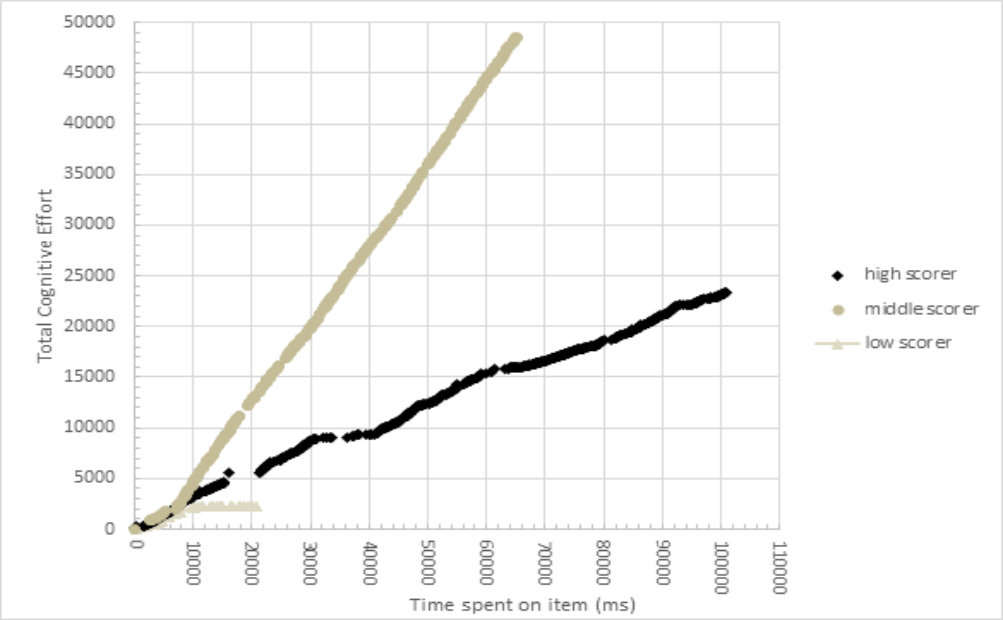

Figure 12. Comparing Total Cognitive Effirt (TCE) for High, Middle and Low Scorer on a Level 6 item.

The cumulative graphs of cognitive effort like Figures 7, 11 and 12 provide evidence related to both the KSA-mitigated model and the assessment construct. Steeper slopes indicate increases in instantaneous effort. These could be co-mapped to sequence maps to see how the increased cognitive load relates to the overall gaze pattern to provide further validity evidence for skill descriptions as well as looking at both optimal solutions and sub-optimal solutions. Horizontal or nearly horizontal areas on these graphs indicate time periods of little cognitive effort beyond that required to focus on a calibration point on the screen. These give evidence of a lack of effortful behaviour that could not be measured simply by latency and click stream data. They also allow for comparing cognitive effort of the same individual across multiple tasks to compare to the complexity metric of the construct.

Finally, differences in these plots could provide insight into test takers' approaches. For example, the Middle Scorer in Figures 11 and 12 has a steep slope for nearly all items while the High Scorer has a much lower slope on all items. Follow-up research and post-test questions about approaches to assessment might provide further insight. Does the Middle Scorer feel the need to rush because of earlier testing results on different assessments? This would validate portions of the CENTS model related to speededness as a contributing factor to total cognitive effort necessary. Is this a case of the proverbial tortoise and the hare? More research is needed to examine these differences.

6.2 Research Question 2: Does the TCE Increase for an Individual as the Anticipated Total Cognitive Difficulty of a Task as Defined by the Construct Increases?

In general, the relationship appeared to be true. Aberrations from this trend may indicate giving up or guessing behaviour if the TCE was far below the expected value. If the value was far above the expected value, it may indicate that the individual was solving using a less than optimal set of KSAs which could be addressed through instruction as suggested by CLT (Paas and Van Merrienboer, 1994; Paas *et al.*, 2003) or there may be facets of the task that caused a greater cognitive load such as the necessity to keep large amounts of information in

working memory because of the density of information or scrolling effects. Figure 12 shows that for individuals with moderate or high graphic literacy, the expected trend in TCE is present and that for individuals with low levels of graphic literacy, the POT was likely exceeded and guessing behaviour occurred after an initial attempt to solve the task.

Figure 8 provides good evidence for both this assertion and the KSA-mitigated model, specifically the Personal Overload Threshold. When looking at Middle and High Scorers, the mean TCE by graphic type increases for all participants except participants three and five who show the expected pattern until COMPLEX graphics are used. Both of these showed low TCE for the items about a COMPLEX process diagram. So, it appeared that they evaluated what they knew and could do and used non-solution behaviours after assessing the graphic stimulus and question stem. This type of behaviour will be discussed further when examining the three students in Figures 11 and 12.

As anticipated the tasks associated with the EASY graphic displayed near automaticity for most of these scorers. For Low scorers, across all three parts of the construct, graphic complexity, number of cognitive steps and overall item level, it appears that participants worked on some problems, particularly those with very familiar graphic types like tables and then decided that the task exceeded their POT and moved on. Figure 9 shows that the Low Scorers exerted approximately twice as much TCE on the item set that focused on two dense tables, but still averaged significantly less TCE than either the Medium or High Scorer groups. This was most evident in the three cognitive step items in the set that had the lowest TCE of the set for all Low Scorers. It appeared that these test takers, in general, know what they can do and do not invest much time or mental effort on tasks that they were not likely to answer correctly which again supports the POT portion of the KSA-mitigated model.

Figure 10 shows the anticipated increases for both High and Middle Scorers. The small sample size makes it difficult to draw conclusions about the groups as some test takers appeared to check their work before moving on which would add to that individual's TCE while others would select an answer and then move on. Also, Figure 12 presents some differences between the Middle and High Scorer. Using latency data alone, it would appear that the Middle Scorer is more efficient, completing the task in approximately 40 seconds less time than the High Scorer. However, the TCE data tells a different story. The Middle Scorer exerted nearly twice as much cognitive effort to complete the same task. Perhaps, the High Scorer was able to mitigate much of the higher instantaneous load by using KSAs to mitigate the overall effort as suggested not only by the KSA-mitigated load model presented here but also by well-known models such as chunking the information (Chi *et al.,* 1981). Other analysis of the data found that individuals with high levels of graphic literacy looked at graphic data and saw trends with automaticity while those with lower graphic literacy skills looked at individual points and then determined if there was a pattern using both heat maps (Thomas and Langenfeld, 2017b; Langenfeld, Thomas, and Gao, 2019) and thinkalouds (Thomas and Langenfeld, 2017a; ACT, 2019) separately.

6.3 Research Question 3: Does the TCE Show Trends that Would Indicate that for Participants who have Low KSAs for a Given Task a Tendency to Cognitively Disengage and Guess?

Yes, response data indicated that when individuals attempt tasks far above their mastery level, as indicated by overall assessment score, TCE generally decreased. For difficult items that were based on more familiar contexts, the relevant KSAs allowed low-scoring individuals to attempt the item because they had not exceeded the POT. However, on most complex items, the individual lacked the necessary KSAs and the POT was exceeded, resulting in guessing. Moreover, for items beyond the simplest categorization of each facet, low scorers showed a decrease in TCE indicating that the items exceeded their abilities and that guessing behaviours occurred.

Additionally, Figures 11 and 12 give evidence of what would be expected. The Level 4 item in Figure 11 is near the ability level of the Low Scorer. The individual exerts a great deal of TCE, over 21,000 units across almost 55 seconds of effort. There are several changes in slope indicating increased processing activity at around 25 and 45 seconds elapsed, with a lower intensity effort section between. These correspond to decreases in mental activity

as measured by pupillometry when changing tasks in previous research (Bailey and Iqbal, 2008). The individual has committed to answering the question suggesting that that the POT and KSAs were sufficient to overcome MIC-demand. Even though the selected answer was incorrect, the individual clearly engaged in meaningful effort.

Conversely, the pattern in Figure 12 for the Level 6 item showed the opposite. The initial nine seconds of reading the question and scanning the graph were followed by a period of virtually no cognitive effort. It is as if the test taker read the question looked at the passage and then posed the mental question 'can I answer this' and decided no and guessed correctly. Using latency data alone would have suggested that this individual engaged in meaningful solution behaviour. The time spent on the question exceeded the established guessing threshold of 15 seconds. It also exceeded the ten percentage of average time threshold suggested by Wise (2017). However, the TCE graph illustrates the principle in the KSA mitigated model addressing POT and non-solution behaviours. The individual got to a point that exceeded POT and then guessed.

6.4 Research Question 4: Does Individual Random Variation in TCE Render Measures of Central Tendency Uninterpretable?

No, individual random variation in TCE measures did not render the measures of central tendency uninterpretable, as shown in Tables 3 and 4. For the Graphic Literacy assessment, individual level TCE data showed weaker correlations to the construct components, between 0.20 and 0.26. However, when using the median TCE values, the correlations to the three constructs components as well as to the IRT a-parameter and b-parameter showed strong relationships. Correlations of the median item TCE to three constructs and item difficulty and discrimination ranged from 0.55 and 0.62. (Thomas and Langenfeld, 2018; Langenfeld *et al.,* 2020).

One of the great limitations of eye-tracking research is the small sample size. In this study, data from 18 individuals was used. However, since each individual generated thousands of data points, there were sufficient data to warrant conclusions. Ideally, the study would be replicated with a larger, more representative sample of middle level scorers. The inclusion of the "Messy Middle" (Gotwals and Songer, 2009; Gotwals, Songer, and Bullard, 2012) could provide additional insight. Since the Messy Middle inconsistently possesses and applies KSAs, they are likely to show greater TCE variability. They might also provide insights into how items just above the Zone of Proximal Development (Vygotsky, 1987) behave for those with incomplete skills as Figure 8 showed for a Low

Table 3. Correlations of individual TCE and mean TCE to assessment constructs

Construct Component	Individual TCE by Item	Median TCE by Item
Graphic Complexity	0.20	0.55
Cognitive Steps	0.26	0.56
Overall Item Level	0.26	0.62

Table 4. Correlations of IRT 3PL parameters to Median TCE by item

IRT Parameter	Correlation to Median TCE by item
a (discrimination)	0.46
b (difficulty)	0.45

Scorer. This could provide additional insight into the POT. Additional data through retrospective think aloud protocols (Someren *et al.*, 1994) could provide validation that the individual was trying to complete the task and at some point reached a sub-task that was too difficult and then proceeded to guess (either on that component of the task or the entire task) and then moved on either to complete the remaining tasks (if the individual only guessed on the sub-component) or to the next item.

6.5 Limitations and Next Steps

The largest limitation to this research is the small sample sizes that can be collected using eye-tracking equipment. The other portion of this study which used paper and pencil also included approximately 20 people; however, the SMI-goggles used are not compatible with the ICA software. This approach should be applied to other assessments and constructs. Data has been collected on a workforce reading and a workforce math assessment but these data have not yet been analyzed.

Since most assessed constructs include a construct map (Wilson, 2009) that suggests that task becomes more cognitively difficult as the items and tasks move along the progression, there should be a method to validate this increased difficulty by using some measure of cognition such as TCE. This approach of integrating the ICA over task time should be applied to other assessments including those based on PNP (Paper and Pencil) or CBT (Computer-based Testing) formats to determine if the findings here are representative of a general pattern that affirms the interaction of KSAs with the cognitive demands of the tasks involved. Eye tracking data have been collected for both math and reading in the workplace assessments that will be analyzed to address this issue. The potential of using cognitive effort as a measure of item difficulty in addition to psychometric indices could lead to a more cognitively-based understanding of item characteristics. For many years, psychometric theory has assumed that the difficulty of an item can be determined by statistical properties whether using classical or item response theory. Psycho-physiological data from eye tracking can now provide insight into both the instantaneous difficulty of parts of a task using cognitive load and the overall difficulty of the task using total cognitive effort. Analysis of these types of data can provide insights which allow for interpretations of observable behaviour that relate to the underlying KSAs and cognitive processes. This analysis should allow researchers to link observation, cognition, and interpretation vertices of the assessment triangle (NRC, 2001).

Additionally, ICA could be used to compare the cognitive effort required for PNP and CBT formats. This would provide evidence to support or refute the claim that the two testing formats can be used interchangeably as ways of assessing the same constructs. For example, does scrolling on a computer screen add cognitive load and effort that is not present when working on a paper test? Does having to scan for information across facing pages on paper require more TCE than the corresponding computer-based testing skill? Are there differences in groups being assessed based on age, access to computers, or other factors that might affect validity arguments?

Subsequent eye-tracking studies included follow-up questions pertaining to how the participant completed the task relative to normal class work. The follow-up study on another assessment included questions about metacognition and the perceived need to rush. Perhaps, differences in high ICA and low ICA may be explained by student perceptions about the need to rush which would validate speedness requirements as contributing to the TCERT.

Integrating psycho-physiological data like pupillometry with other sources of validity evidence including both traditional cognitive labs and psychometric item properties should strengthen arguments about the validity of uses and claims based on an assessment. For example, tasks that require three-dimensional science reasoning as defined by the Next Generation Science Standards (NGSS Lead States, 2013) should show greater TCE than tasks that only require a student to engage in one or two of the dimensions. Since 2001 when Knowing What Students Know (NRC, 2001) was published, there have been efforts to get inside the black box of cognition. This KSA-mitigated model and use of pupillometry to find total cognitive effort for tasks present a step towards making that unseeable portion of the assessment triangle measurable.

7. References

Aberg-Bengtsson, L., Ottosson, T. (2006). What lies behind graphicacy? Relating students' results on a test of graphically represented quantitative information to formal academic achievement. *Journal of Research in Science Teaching*. 43(1):43–62. https://doi.org/10.1002/tea.20087.

ACT (2019).WorkKeys Graphic Literacy technical manual. March 21, 2022. https://www.act.org/content/dam/act/unsecured/documents/WorkKeys-Graphic-Literacy-Technical-Manual.pdf.

American Educational Research Association (AERA), American Psychological Association (APA) and National Council for Educational Measurement (NCME). (2014). Standards for educational and psychological testing (2014 ed.). Washington, DC: AERA Publications.

Aminihajibashi, S., Hagen, T., Foldal, M. D., Laeng, B., Espeseth, T. (2019). Individual differences in resting-state pupil size: Evidence for association between working memory capacity and pupil size variability. *International Journal of Psychophysiology*. 140:1–7. PMid: 30894328. https://doi.org/10.1016/j.ijpsycho.2019.03.007.

Ahern, S., Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science*. 205(4412):1289–92. PMid: 472746. https://doi.org/10.1126/science.472746.

Andrzejewska, M., Stolinska, A. (2016). Comparing the difficulty of tasks using eye tracking combined with subjective and behavioural criteria. *Journal of Eye Movement Research*. 9(3). https://doi.org/10.16910/jemr.9.3.3.

Ayaz, H., Shewokis, P. A., Bunce, S., Onaral, B. (2011, August). An optical brain computer interface for environmental control. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 6327–30). *IEEE*. PMid: 22255785. https://doi.org/10.1109/IEMBS.2011.6091561.

Bailey, B. P., Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*. 14(4):1–28. https://doi.org/10.1145/1314683.1314689.

Bartels, M., Marshall, S.P. (2006, March). Eye tracking insights into cognitive modeling. In Proceedings of the 2006 symposium on Eye Tracking Research and Applications (pp. 141–7). https://doi.org/10.1145/1117309.1117358.

Beatty, J. (1982). Task-evoked pupillary responses, processing load and the structure of processing resources. *Psychological Bulletin*. 91:276–92. PMid: 7071262. https://doi.org/10.1037/0033-2909.91.2.276.

Beatty, J., Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*. 5(10):371–2. https://doi.org/10.3758/BF03328444.

Beatty, J. Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary and G. G. Berntson (Eds), Handbook of Psychophysiology (2nd ed.; pp. 142–62). Cambridge, UK: Cambridge University Press.

Beauchamp, M. R. (2016). Disentangling motivation from self-efficacy: Implications for measurement, theory-development and intervention. *Health Psychology Review*. 10(2):129–32. PMid: 26953186. https://doi.org/10.1080/17437199.2016.1162666.

Blessing, S. B., Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 22(3):792–810. https://doi.org/10.1037/0278-7393.22.3.792.

Boehm-Davis, D. A., Gray, W. D., Adelman, L., Marshall, S., Pozos, R. (2003). Understanding and measuring cognitive workload: A coordinated multidisciplinary approach. George Mason University Fairfox VA Department of Psychology. Retrieved March 22, 2022. https://apps.dtic.mil/sti/pdfs/ADA417743.pdf https://doi.org/10.21236/ADA417743.

Bryant, P. E., Somerville, S. (1986). The spatial demands of graphs. *British Journal of Psychology*. 77:187–97. PMid: 3730726. https://doi.org/10.1111/j.2044-8295.1986.tb01993.x.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., Conway, D. (2016). Robust multimodal cognitive load measurement. Switzerland: *Springer International Publishing*; https://doi.org/10.1007/978-3-319-31700-7.

Chi, M. T., Feltovich, P. J., Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*. 5(2):121–52. https://doi.org/10.1207/s15516709cog0502_2.

Curcio, F.R. (1987). Comprehension of mathematical relationships expressed in graph. *Journal for Research in Mathematics Education*. 18(5):382–93. https://doi.org/10.5951/jresematheduc.18.5.0382.

Curcio, F.R., Artz, A.F. (1997). Assessing students' statistical problem solving behaviours in small-group setting. In Gal, I., and Garfield, J.B. (Eds.). The assessment challenge in statistics education Amsterdam: IOS Press; p. 123–38.

Ericsson, K. A., Simon, H. A. (1980). Verbal reports as data. *Psychological Review*. 87(3):215–50. https://doi.org/10.1037/0033-295X.87.3.215.

EyeTracking, Inc. (2014). Workload Module Manual.

Friel, S. N., Bright, G. W. (1996). Building a theory of graphicacy: How do students read graphs? Retrieved September 30, 2021. http://eric.ed.gov/?id=ED395277.

Friel, S. N., Curcio, F. R., Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*. 32(2):124–58. https://doi.org/10.2307/749671.

Garner, R. (1987). Metacognition and reading comprehension. Norwood, NJ: Ablex Publishing.

Gotwals, A. W., Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*. 94(2):259–81. https://doi.org/10.1002/sce.20368.

Gotwals, A. W., Songer, N. B., Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. A.C. Alonzo and A.W. Gotwals (Eds.). *Learning progressions in science*. Rotterdam: Sense Publishers; p. 183–210. https://doi.org/10.1007/978-94-6091-824-7_9.

Gutirrez, R. S., Shapiro, L. P. (2010). Measuring the time course of sentence processing with pupillometry. Cuny conference on human sentence processing.

Hess, E. H., Polt, J. R. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*. 143(3611):1190–2. https://doi.org/10.1037/0033-295X.87.3.215.

Hmelo-Silver, C. E., Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors and functions. *Cognitive Science*. 28(1):127–38. https://doi.org/10.1207/s15516709cog2801_7.

Just, M.A., Carpenter, P.A., Woolley, J.D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology*. 111(2):228–38. PMid: 6213735. https://doi.org/10.1037/0096-3445.111.2.228.

Kahneman, D. (2011). Thinking, fast and slow. New York, NY: Farrar, Straus, and Giroux;.

Kahneman, D., Beatty, J. (1966). Pupil diameter and load on memory. *Science*. 154(3756):1583–5. PMid: 5924930. https://doi.org/10.1126/science.154.3756.1583.

Kliewer, K., Langenfeld, T., Thomas, J. (2018, October). Observing cognitive processes and identifying reading difficulties through eyetracking analysis. Poster session presented at the Education Technology and Computational Psychometrics Symposium in Iowa City, IA.

Langenfeld, T., Thomas, J., and Gao, X. (2019, April). Principled assessment design: Applications and tools for assessment updates. Paper presented at the meeting of the National Council for Educational Measurement, Toronto, Ontario.

Langenfeld, T., Thomas, J., Zhu, R., Morris, C. A. (2020). Integrating multiple sources of validity evidence for an assessment-based cognitive model. *Journal of Educational Measurement*. 57(2):159–84. https://doi.org/10.1111/jedm.12245.

Lee, S. C., Nathan, B. R. (1997). Obtaining reliable job analysis information: A progress report from the Work Keys System TM: ACT's nationwide program for improving workplace skills. March 22, 2022. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.679&rep=rep1&type=pdf.

Manseta, K., Khwaja, A. M., Sultan, E., Daruwalla, P., Pourrezaei, K., Najafizadeh, L., Gandjbakhche, A., Daryoush, A. S. (2011, June). Untethered helmet mounted Functional Near Infrared (fNIR) biomedical imaging? Microwave Symposium Digest (MTT), 2011 IEEE MTT-S International (pp. 1-4). *IEEE*. https://doi.org/10.1109/MWSYM.2011.5972982.

Marshall, S. P. (2000). U.S. Patent No. 6,090,051. Washington, DC: U.S. Patent and Trademark Office.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. Human factors and power plants, 2002, proceedings of the 2002 IEEE 7th Conference. Scottsdale, AZ: *IEEE*. p. 75–9. https://doi.org/10.1109/HFPP.2002.1042860.

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation, Space and Environmental Medicine*. 78(5Sec II, Suppl):B165–75.

Marshall, S. P., Davis, C. L., Knust, S. R. (2004). The index of cognitive activity: Estimating cognitive effort from pupil dilation. San Diego, CA: EyeTracking, Inc.; Technical Report ETI-0401.

Marshall, S. P., Pleydell-Pearce, C. W., Dickson, B. T. (2003). Integrating psycho-physiological measures of cognitive workload and eye movements to detect strategy shifts. Proceedings of the 36th Annual Hawaii International Conference on System Sciences. https://doi.org/10.1109/HICSS.2003.1174298.

Martinez, M. E., Katz, I. R. (1995). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational*

Assessment. 3(1):83–98. https://doi.org/10.1207/s15326977ea0301_4.

Morrison, J. G., Marshall, S. P., Kelly, R. T., Moore, R. A. (1997). Eye tracking in tactical decision making environments: Implications for decision support evaluation. Third International Command and Control Research and Technology Symposium, National Defense University, June 17-20, 1997. May 14, 2022. http://www.all.net/journal/deception/www-tadmus.spawar.navy.mil/www-tadmus.spawar.navy.mil/Eye_Trkr.pdf.

National Research Council (2001). Knowing what students know: The science and design of educational assessment. National Academies Press.

NGSS Lead States (2013). Next Generation Science Standards: For states, by states. Washington, DC: The National Academies Press.

Paas, F. G. W. C, Van Merrienboer J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*. 6:351–71. https://doi.org/10.1007/BF02213420.

Paas, F., Tuovinen, J. E., Tabbers, H., Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*. 38(1):63–71. https://doi.org/10.1007/BF02213420.

Palmer, H., Valet, W. (2001). Job analysis: Targeting needed skills. *Employment Relations Today*. 28(3):85–92. https://doi.org/10.1002/ert.1029.

Pomplun, M., Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. Proceedings of the 10th International Conference on Human Computer Interaction, V.D.D. Harris, M. Smith, and C. Stephanidis, Eds.

Poole, A., Ball, L. J. (2006). Eye tracking in HCI and usability research. C. Ghaoui (Ed.) Encyclopedia of Human-Computer Interaction. *Pennsylvania: Idea Group*; p. 211–9. PMid16524094. https://doi.org/10.4018/978-1-59140-562-7.ch034.

Schiff, M. (1875). La pupille considérée comme esthésiomètre. Baillière.

Sensori Motoric Instruments (SMI). (2016). iView User Guide: Version 2.7.1. Boston, MA: Author.

Shah, P., Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*. 3(3):560–78. PMid: 25164403. https://doi.org/10.1111/j.1756-8765.2009.01066.x.

Shah, P., Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*. 14(1):47–69. https://doi.org/10.1023/A:1013180410169.

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

Tan, J. K., Benbasat, I. (1990). Processing of graphical information: A decomposition taxonomy to match data extraction tasks and graphical representations. *Information Systems Research*. 1(4):416–39. https://www.jstor.org/stable/23010666.

Tanaka, Y., Yamaoka, K. (1993). Blink activity and task difficulty. *Perceptual and Motor Skills*. 77(1):55–66. PMid: 8367265. https://doi.org/10.2466/pms.1993.77.1.55.

Taylor, H. A., Renshaw, C. E., Choi, E. J. (2004). The effect of multiple formats on understanding complex visual displays. *Journal of Geoscience Education*. 52(2):115–21. https://doi.org/10.5408/1089-9995-52.2.115.

Thomas, J., Langenfeld, T. L. (2017a, April). Analyzing think-aloud and eye-tracking data to support score interpretations. Paper presented at the meeting of the National Council on Measurement in Education, San Antonio, TX.

Thomas, J., Langenfeld, T. (2017b, November). Using multiple lines of evidence to validate a construct and assessment. Poster presented at the Education Technology and Computational Psychometrics Symposium, Iowa City, IA.

Thomas, J., Langenfeld, T. L. (2018). Using eye tracking and pupillometry to validate a Graphic Literacy assessment. Presentation at ITC 2018, Montreal, Canada.

Thomas, J., Langenfeld, T. (2019, April). Using eye tracking data to validate the cognitive processes of foundational workplace skills. Paper presented at the meeting of the National Council on Measurement in Education, Toronto, Ontario.

Tobii Pro, Inc (2016). Tobii Pro Glasses 2 User's Manual.

Veltman, J. A., Jansen, C. (2006). The role of operator state assessment in adaptive automation. *TNO Defense, Security and Safety*. March 22, 2022. https://apps.dtic.mil/sti/pdfs/ADA455055.pdf.

Vygotsky, L. (1987). Zone of proximal development. Mind in society: The development of higher psychological processes, 5291:157.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*. 21(1):14–23. https://doi.org/10.3102/0013189X021001014.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching: The Official Journal of the*

National Association for Research in Science Teaching. 46(6):716–30. https://doi.org/10.1002/tea.20318.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation and implications. *Educational Measurement: Issues and Practice*. 36(4):52–61. https://doi.org/10.1111/emip.12165.

Wise, S. L., Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. M.J. Margolis and R.A. Feinberg (Eds.). Integrating timing considerations to improve testing practices. p. 150–64. https://doi.org/10.4324/9781351064781-11.

Wise, S. L., DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*. 15(1):27–41. https://doi.org/10.1080/10627191003673216.

Wise, S. L., Pastor, D. A., Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*. 22(2):185–205. https://doi.org/10.1080/08957340902754650.

Wise, S. L., Smith, L. F. (2011). A model of examinee test-taking effort. J. A. Bovaird, K. F. Geisinger, amd C. W. Buckendahl (Eds.). High-stakes testing in education: Science and practice in K-12 settings. *American Psychological Association*. p. 139–53. https://doi.org/10.1037/12330-009.

Young, M. S., Stanton, N. A. (2001). Mental workload: Theory, measurement and application. Karwowski, W. (Ed.). *International Encyclopedia of Ergonomics and Human Factors*. London: Taylor and Francis; Vol. 1. p. 507–9.