

Accounting for Affective States in Response Processing Data: Impact for Validation

Jacqueline P. Leighton

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta


Preparation of this conference paper was supported by a grant to the author from the Social Sciences and Humanities Research Council of Canada (SSHRC Grant No. 410-2011-0811). Grantees undertaking such projects are encouraged to express freely their professional judgment. This paper, therefore, does not necessarily represent the positions or the policies of the Canadian government, and no official endorsement should be inferred. Correspondence pertaining to this article should be directed to the author (jacqueline.leighton@ualberta.ca).

Abstract

The Standards (AERA, APA, NCME, 2014) are clear about the importance of response process data as a source of validity evidence for test and item score interpretation. The purpose of this paper is to summarize how affective/emotional processes influence test performance, then explain and propose reasons to include research on affective processes in validation arguments.
Accounting for Affective States in Response Processing Data:

Impact for Validation

The Standards (AERA, APA, NCME, 2014) are clear about the importance of response process data as a source of validity evidence for test and item score interpretations. However, despite 15 years of attempts by psychometric practitioners and researchers to better integrate response processes into validation efforts, key variables have not been fully considered for gathering and interpreting these data. In particular, one class of variables that has remained virtually unexplored is examinees’ affective or emotional states and how these states might interrupt not only the process of response articulation (i.e., expressing problem solving processes) but also the actual problem solving involved (Leighton, 2013; Norris, 1990; see also Hsu, Babeva, Feng, Hummer & Davison, 2014, for effect of distractions on response processing data).

Understanding the relationship between examinees’ affective or emotional states and their response process data is necessary to ensure these data accurately inform the objectives of test validation. For example, consider that examinees, who were led to believe an expert in the subject matter was interviewing them in a think-aloud study, produced fewer correct responses to test items and less sophisticated cognitive processes (models) compared to examinees who thought a novice was conducting the interview (Leighton, 2013). The purpose of this paper is to (a) identify the relevance of emotions and, in particular, evaluative (test) anxiety in academic test performance and response process data as a source of validity evidence, (b) summarize recent research suggesting a possible link between evaluative anxiety and response process data collected in think aloud studies, and (c) call for research to address affective or emotional states in the collection and interpretation of response process data for the purpose of validation.
Relevance of Academic Emotions:
Affect, Evaluative Anxiety, Tests and Validation

According to the Merriam-Webster Dictionary, the term ‘affect’ is a feeling in the first sense and the conscious subjective aspect of an emotion considered apart from bodily change in the second sense. Affect and emotion are similar terms but ‘emotion’ is broader in meaning, including the unconscious experience of a feeling, often leading to physiological changes. For example, anxiety can manifest itself consciously with concomitant physiological changes such as increased heart rate, sweat gland and bladder activity. Anxiety can also manifest itself unconsciously with an elevated neuroendocrine response consisting of increased epinephrine, norepinephrine, cortisol, growth hormone and prolactin; the latter physiological responses have been noted in response to a variety of stressful circumstances including engaging in extreme sports, participating in benign laboratory conditions, and completing academic examinations (Hoehn-Saric & McLeod, 2000). For the purposes of this paper, the terms affect and emotion are used interchangeably to mean the experience of feeling, which may include physiological changes in the body.

The range of emotions students experience in school settings related to their learning and achievement has only recently began to receive the full attention it deserves (Pekrun, Goetz, Titz, & Perry, 2002). Over a decade ago, a group of prominent educational psychologists, Pekrun et al. (2002, p. 91) wrote “[A]cademic emotions have largely been neglected by educational psychology, with the exception of test anxiety.” For example, hope, pride, relief, anger, hopelessness, shame and boredom were previously neglected and yet often observed, anecdotally, to influence student learning and achievement on learning tasks, in classroom participation and achievement testing. To be sure, since 2002, this research deficit has been systematically addressed and as an example, the first International Handbook of Emotions in Education was
Affective States in Response Processes

released in 2014, including chapters on anxiety, interest, enjoyment, boredom, shame and pride (see Reinhard & Linnenbrink-Garcia, 2014). Moreover, affective states including trust and wellbeing are also increasingly the focus of study in educational measurement (e.g., Chu, Guo & Leighton, 2014).

Despite a newfound focus on academic emotions generally, evaluative anxiety, specifically, has long been known to impede student test performance and continues to be a well-established impediment (Cizek & Burg, 2006; Zeidner, 2014). The Yerkes-Dodson law (i.e., the inverted U-shaped curve showing the relationship between performance and arousal) reminds us that mild forms of arousal can be beneficial to performance by focusing attention on the task of interest. However, arousal and anxiety are distinct; arousal is the state of being “alert” or “awake,” whereas anxiety is a negative physiological state, characterized by nervous behaviours and thoughts.

Many students experience anxiety in their academic lives. In fact, according to the American Test Anxieties Association, anxiety is considered to be the dominant source of scholastic impairment with a prevalence rate of 16-20% of students reporting high test anxiety and another 18% reporting moderate anxiety. Aside from recognizing the harmful effects of anxiety on test performance and trying to provide tips and strategies for reducing it (Cizek & Burg, 2006), there are no formal ways to treat it except to provide test-taking accommodations for students. Usually these come in the form of more testing time. However, to be considered for accommodations under the Americans with Disabilities Act (ADA), students may need to demonstrate evidence of “mental impairment” and show that it substantially limits one or more major life activities (Zuriff, 1997). Having to provide such evidence is not without its social drawbacks.

Anxiety is often most pronounced during performance-oriented school activities such as
test taking or in anticipation of taking a test. That test items can provoke anxiety and interfere with examinees’ response processing is well established (Cizek & Burg, 2006) and unsurprising given that achievement tests, including other tests such as college-readiness tests, are often high-stakes and gateways to postsecondary opportunities and economic mobility. Test items, which frequently rely on a multiple-choice format and are heavily focused on performance, often arouse concern and even alarm in examinees (Ryan & Ryan, 2005). For example, many girls and minority examinees exhibit test anxiety in fear of fulfilling negative stereotypes when completing multiple-choice mathematics tests (Bosson, Haymovitz, & Pinel, 2004; Cohen & Sherman, 2005; Steele, 1997). Participant sensitivity to the features or conditions of a test or test environment falls under the umbrella of demand characteristics or subject-experimenter effects (Christensen, 1991) and need to be considered when collecting response process data.

Even if students with debilitating high anxiety are accommodated, many more undoubtedly experience moderate anxiety, do not report it, and are not accommodated. But does it disturb their test performance nonetheless? Making inferences about non-accommodated students’ achievement test scores in light of moderate anxiety impairments is bound to be challenging given that no formal or systematic remedies exist for recognizing or controlling such impairments. Test or evaluative anxiety is usually not taken into account in the interpretation of test scores or in validation efforts with non-accommodated populations; probably due to the assumption that if it is moderate, it can be controlled or it does not alter total test performance in a material way. However, in validation studies, where the objective is to drill down on individual test items and collect finer-grained data about response processing, moderate anxiety may matter. Moderate anxiety may alter the strategies used to solve items and influence what test developers conclude about what those items are measuring. Surprisingly, few if any published papers or reports outline why and how evaluative anxiety might influence response process data gathered in
think-aloud interviews – one of the primary methodologies used to collect response process data (Leighton, 2004). However, there is reason to be concerned and need to consider taking into account students’ emotions, in particular, evaluative anxiety.

**Emotion and Disruptions of Cognitive Processing:**

**Implications for Validation**

Response process data are increasingly collected using think aloud interviews as a source of validity evidence for test items (Messick, 1994; see also Leighton, 2004). However, given the potentially negative role of evaluative anxiety in academic test performance (e.g., Cassady & Johnson, 2002; Zeidner, 2014), it begs the question of how response process data may be compromised during think aloud interviews; especially validation efforts that involve students engaging with the very test materials that often arouse anxiety. Over the past 25 years, think-aloud interview studies have been increasingly used to gather data from students as they respond to test items. The collection of data to empirically verify the response processes students use to solve test items is strongly recommended by many educational testing professionals (Leighton, 2004). The data collected informs not only test item comprehension, design and development (Leighton & Gierl, 2007), but also validation of the expected knowledge and skills being tested and whether these are indeed used by examinees to solve the items. Furthermore, the response processes used by students of different ability levels to solve the items can be used to confirm expectations about the cognitive strategies underwriting correct and incorrect responses.

However, just as evaluative anxiety may disrupt performance in actual testing situations (Cizek & Burg, 2006), it is likely that evaluative anxiety – even moderate, self-regulated anxiety - disrupts response processes in think-aloud interviews.

Anxiety – even moderate anxiety – may disrupt response processing of test items during think-aloud studies. First, students participating in think-aloud interviews may perceive the
interviews to be evaluative of their abilities, become self-conscious, and experience negative affect. To be sure, even if the interviews are presented by investigators as low-stakes, students may nonetheless view them as high-stakes because a third-party is watching them problem solve (see Kyllonen, in press; Sawyer Jr. & Hollis-Sawyer, 2005; Steele, 1997). In other words, simply engaging with test items in front of the investigator or interviewer may elevate stress levels for students. Second, and related to the first, assuming that students perceive the think-aloud interview as another testing opportunity, evaluative anxiety may impact performance in much the same way as it would under usual testing circumstances.

The cognitive mechanism by which anxiety is expected to influence response processing involves impairment of working memory or its central executive. For example, a dominant view has been that the anxiety experienced during evaluative situations leads to excessive self-monitoring of performance, which overburdens working memory and leaves scarce attentional resources to focus on the task at hand (see Sarason, Sarason & Pierce, 1995). For example, Beilock, Kulp, Holt, and Carr (2004) identified a state of “choking under pressure” in situations where performance is considered to be highly valued and potentially used to make inferences about individual achievement or intellect (i.e., ego-threatening). Students in these performance-oriented situations have been found to engage in excessive self-monitoring, the result of which is disrupted response processing and suboptimal outcomes (see also Beilock & Carr, 2001; Ericsson, 2006; Lewis & Linder, 1997). In general, excessive self-focus reduces attention on the task and impairs performance.

Another more recent mechanism is outlined in attention-control theory (Eysenck, Derakshan, Santos, & Calvo, 2007). In attention-control theory, anxiety is shown to disrupt performance not necessarily because of excessive self-monitoring but because the individual is constantly engaging in an outward focused, environmental scan, looking for potential threats, for
example, the facial expressions of the interviewer; thus hampering the central executive of working memory to regulate and direct needed attentional resources to focus on the task. Neuroscientific evidence indicates that anxiety increases amygdala activation and decreases activation of prefrontal cortical areas, which are known to be extensively involved in the regulation of attention (Bishop, 2007). A recent study by Hsu et al. (2014) experimentally investigated induced distractions (i.e., answering trivia questions, playing a visual puzzle game [tetris], or no distraction [control]) during a think-aloud study, and found that the distractions significantly altered cognitive processing but not emotional states, for example, anger about having to think aloud. In particular, a content analysis of the verbal reports using the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) revealed students in one of the distraction conditions – answering trivia questions and playing tetris – resulted in a lower number of words during the think aloud session relative to the control. In addition, answering trivia questions led to the production of more non-fluencies (e.g., “uh” or “ummm”) and filler words (e.g., “like” or “you know”) not related to the task and greater measured disengagement compared to controls.

Evaluative anxiety may have implications for the accuracy of response process data gathered during think-aloud interviews, and then used for test validation efforts. Unfortunately, however, relatively little research has focused on measuring the impact of evaluative anxiety on response processing during think aloud studies. One study, conducted by Leighton (2013) is illustrative of the reason for concern over potential consequences. In a think-aloud study of 15 items sampled from a standardized, Grade 12 exit-test in mathematics, Leighton (2013) found students’ response processes altered as a function of interviewer characteristics. Leighton randomly assigned 71 high and moderate-ability Grade 12 students to one of three think-aloud conditions. Each of the conditions was exactly the same in terms of think-aloud task instructions,
order of test items, and order of follow-up surveys, except for one variable – the script used for how the interviewer was introduced. In the first condition, called the *novice interviewer condition*, the interviewer introduced himself/herself as a non-expert in mathematics using the following script:

> Thank you for taking part in today’s study. Now, before I explain what we will be doing today, let me introduce myself. My name is _________ and I’m from the University of _____.* My area of expertise is not in *Mathematics but* I’ve been interested in how students solve problems for many years. So, now let me tell you about the study you’re involved with today. I will read this because it is important and I want to make sure everyone gets the same instructions….

In the second condition, called the *expert interviewer condition*, the interviewer introduced himself/herself as an expert in mathematics and stated *My area of expertise is in Mathematics* – the script was otherwise identical to the non-expert condition. In the third condition (control), the interviewer did not include any mention of his or her expertise in mathematics.

Leighton (2013) found that students assigned to the novice interviewer condition were more accurate in their responses for easy, medium and difficult mathematics items compared to students assigned to the control and expert conditions. Furthermore, students assigned to the novice interviewer condition exhibited more sophisticated responses processes (as coded in retrospective verbal reports) on medium and difficult items compared to students in both the expert/control conditions. Although Leighton found no differences across the three conditions on the *Test Attitude Inventory*, a self-report measure of state test anxiety, and on measures of familiarity or confidence, and even metacognition (self-regulation), differences were found in how students perceived the interviewer. Students in both the expert and control condition
reported perceiving the interviewer “as an expert in Mathematics” significantly more often than those in the novice condition. Furthermore, on an indirect, measure of nervousness – frequency of validation seeking speech (e.g., am I on the right path?) – students in the expert and control conditions tended to exhibit more of this speech compared to students in the novice condition although this difference failed to reach statistical significance (see also Hsu et al., 2014 for similar findings).

Closer scrutiny needs to be paid to the think-aloud interview conditions for collecting response process data. A limitation with the Leighton (2013) study is that self-report measures of state anxiety may be insufficiently sensitive to detect the emotion of interest (see Kyllonen, in press). The indirect speech-based measure of nervousness may be more promising because it is performance-based (see Kyllonen, in press). However, this measure may require a larger sample size to have sufficient power to detect changes across distinct treatment conditions. Nonetheless, the case remains that response process data, gathered using traditional think-aloud methods, may not provide a clear window into students’ thinking as they respond to test items. Although the most likely explanation for the reduced accuracy and response processing across interviewer conditions in the Leighton (2013) study is evaluative anxiety leading to disruption of student’s attentional focus on test items, more research is needed. However, given previous research on how evaluative anxiety influences test performance and the effects of third party observers (Kyllonen, in press), it is difficult to imagine a different account for why students’ accuracy and response processes declined as a result of interviewers’ introduction of themselves.

The potential imprecision associated with response process data gathered from think-aloud interviews has implications for validation generally and validity arguments in particular. As response process data are increasingly used to confirm the type, range and sophistication of response processes examinees use to construct and/or select answers to test items (see AERA,
APA, & NCME, 2014), we must be vigilant that the data do indeed provide helpful, clarifying information about students’ response processes. At the very least, what we observe in the Leighton (2013) study is that response process data may be easily distorted as function of environmental conditions such as how interviewers identify themselves. If such a slight manipulation to the conditions of the think-aloud interview can lead to reduced accuracy and quality of knowledge and skills, imagine what other, seemingly benign variables may be distorting response process data? This may be an especially important consideration in think-aloud studies of differential item functioning, when investigators are looking for how item characteristics may bias or distort examinees’ interpretation and thinking about test items.

A Research Call:

Investigating Emotion and its Effect on Response Process Data

Given the increasing prevalence of employing think aloud interviews to collect response-processing data, it is imperative for researchers to consider the variables that may impact the quality of data generated. Although the Standards (AERA, APA, NCME, 2014) indicate the benefit of collecting response-processing data to bolster validity arguments, very little is still known about the quality of these data, the methodological procedures and interview conditions for generating the highest quality data possible, and best practices for coding, interpreting and integrating these data within developing validity arguments.

Educational measurement specialists must consider ways to investigate the role and potential impact of affect and emotion on the response processing of test items. Studies, similar to Leighton (2013), using a variety of experimental and non-experimental methods with a cross section of student populations need to be conducted, using increasingly more fine-grained analyses. Large-scale research programs need to shed light on important methodological issues related to think-aloud interviews, including identifying (a) the range of emotions that may impact
response processing data, (b) the interview conditions that will defuse evaluative anxiety or other emotions that may impede students’ response processing, (c) the mechanisms by which evaluative anxiety or other emotions distract students and disrupt response processing, and (d) ways to account or control for evaluative anxiety or other stress-inducing emotions if they cannot be eliminated during think-aloud studies. It is notable that in both Leighton (2013) and Hsu et al. (2014), clear expressions of the lack of research in this area are made evident. For example, Hsu et al. (2014, p. 1) state “Although the detrimental effects of distraction on a variety of more basic cognitive tasks (e.g., visuo-spatial and working memory tasks) are well known (e.g., Lavie, 2005; Tremblay et al., 2005), studies explicitly examining the effects of distraction on engagement in cognitive-affective think-aloud paradigms are notably absent.” Educational measurement specialists are in a unique and advantageous position to carry out this research given the expertise and rigor we impose on the measurement and interpretation of student achievement and learning.
References


