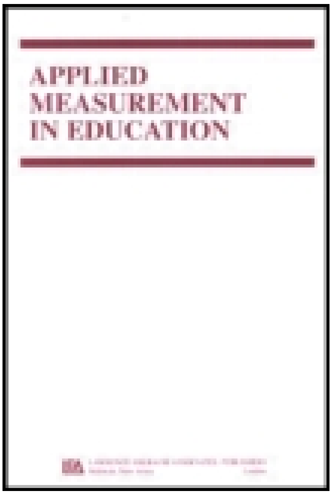

Applied Measurement in Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hame20

Item Difficulty and Interviewer Knowledge Effects on the Accuracy and Consistency of Examinee Response Processes in Verbal Reports

Jacqueline P. Leighton [a]

[a] Centre for Research in Applied Measurement and Evaluation, Faculty of Education , University of Alberta

Published online: 09 Apr 2013.

To cite this article: Jacqueline P. Leighton (2013) Item Difficulty and Interviewer Knowledge Effects on the Accuracy and Consistency of Examinee Response Processes in Verbal Reports, Applied Measurement in Education, 26:2, 136-157, DOI: 10.1080/08957347.2013.765435

To link to this article: http://dx.doi.org/10.1080/08957347.2013.765435

Applied Measurement in Education, 26: 136–157, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 0895-7347 print/1532-4818 online
DOI: 10.1080/08957347.2013.765435

Routledge
Taylor & Francis Group

Item Difficulty and Interviewer Knowledge Effects on the Accuracy and Consistency of Examinee Response Processes in Verbal Reports

Jacqueline P. Leighton
Centre for Research in Applied Measurement and Evaluation, Faculty of Education, University of Alberta

The Standards for Educational and Psychological Testing indicate that multiple sources of validity evidence should be used to support the interpretation of test scores. In the past decade, examinee response processes, as a source of validity evidence, have received increased attention. However, there have been relatively few methodological studies of the accuracy and consistency of examinee response processes as measured by verbal reports in the context of educational measurement. The objective of the current study was to investigate the accuracy and consistency of examinee response processes—as measured by verbal reports—as a function of varying interviewer and item variables in a think aloud interview within an educational measurement context. Results indicate that the accuracy of responses may be undermined when students perceive the interviewer to be an expert in the domain. Further, the consistency of response processes may be undermined when items that are too easy or difficult are used to elicit reports. The implications of these results for conducting think-aloud studies are explored.

With the growing interest in using standardized achievement tests to assess students' mastery of knowledge and skills, methods to validate test score inferences have become a topic of concern for researchers and practitioners. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) indicate that multiple sources of validity evidence can be used to support the interpretation of test scores. Sources of validity evidence can include test content, examinees' response processes, internal item structures, and the relation of test scores to other variables. Of particular interest are examinees' response processes because they are expected to provide evidence of the knowledge and skills examinees use to solve test items. One of the methods used to collect data on examinees' response processes is the *think-aloud interview* (Ericsson & Simon, 1993; see also Ercikan et al., 2010; Leighton, 2004). Although Ericsson and Simon (1993) reviewed a comprehensive list of studies indicating that think-aloud interviews yield accurate and

This research was supported by a grant (#410-2007-1142) from the Social Sciences and Humanities Research Council of Canada (SSHRC).

Correspondence should be addressed to Jacqueline P. Leighton, Ph.D., University of Alberta, Centre for Research in Applied Measurement and Evaluation, Faculty of Education, 6-110 Education North, Edmonton, Alberta T6G 2G5, Canada. E-mail: jacqueline.leighton@ualberta.ca

consistent data (i.e., in the form of verbal reports) about participants' response processes, most of the studies they reviewed involved psychological tasks and not standardized achievement test items.

There is reason to hypothesize that the content of verbal reports obtained from standardized achievement test items may be less accurate and consistent than the content of verbal reports obtained from psychological tasks. The rationale for this hypothesis rests with the observation that standardized achievement test items, unlike many psychological tasks, are known to be performance-oriented, have right and wrong answers, and used to evaluate examinees. Consequently, examinees may feel nervous responding to test items in front of an adult interviewer. This nervousness could influence the accuracy of their responses and the consistency of their processes to solve standardized achievement test items (see Leighton, 2004; Wilson, 1994). Therefore, the objective of the current study is to investigate the accuracy and consistency of examinee response processes, as measured by verbal reports, in a performance-oriented, educational measurement context. As think-aloud interviews are increasingly used to generate validity arguments, it becomes necessary to ensure the fidelity of verbal reports when specific variables are manipulated such as interviewer and item variables. The present article is divided into four main sections. The first section presents the rationale for the hypothesis that verbal reports obtained from think-aloud interviews conducted in an educational measurement context may be susceptible to inaccuracies and inconsistencies due to interviewer and item variables. The second section describes an experimental study conducted to investigate the effects of interviewer and item variables on the accuracy and consistency of examinees' response processes as measured by verbal reports. The third section presents the results obtained from the experimental study conducted. The fourth section concludes with the implications of the results for researchers conducting think-aloud interviews in an educational measurement context.

POSSIBLE THREATS TO VERBAL REPORTS IN AN EDUCATIONAL MEASUREMENT CONTEXT

Verbal Reports

The data collected in think-aloud interviews are called *verbal reports* (see Ericsson & Simon, 1993 for a comprehensive review of techniques; for *cognitive labs* see Desimone & LeFloch, 2004; Willis, 2005; Zucker, Sassman, & Case, 2004). Verbal reports provide an auditory or written record of the response processes used by participants to answer a problem-solving task. Verbal reports consist of two parts. These two parts are gathered using standard, formalized procedures, normally involving a one-on-one interview where a participant is requested to "think aloud" as he or she is solving the problem. In the first part, the participant is asked to express the steps used to solve the problem and the interviewer audio records the contents of this *concurrent* report for later analysis. In the second part, the participant is asked to recall how the task was solved and the interviewer audio records the contents of this *retrospective* report for comparison with the concurrent report. The concurrent and retrospective reports are then compared to evaluate the consistency of response processes used to solve the problem. As a procedural guideline, Ericsson and Simon (1993) recommend that the interviewer not be in a student's line of sight when the interview is being conducted so as to avoid distracting the student. However, researchers often avoid this recommendation to approximate a more natural interview with a student, one that minimizes the artificiality of the verbal report without compromising the quality of the data (however, see procedures for cognitive labs outlined by American Institutes for Research [AIR], 2009).

Conducting think-aloud interviews with school-age students has grown in prevalence over the last two decades as a way to identify the knowledge and skills students use to solve standardized achievement test items (e.g., Ferrara & DeMauro, 2006; Gorin, 2006; Leighton, Cui, & Cor, 2009; Snow & Lohman, 1989). Verbal reports of examinees' response processes are believed to provide evidence for validity arguments about the knowledge and skills measured by achievement tests and the defensibility of test-based inferences (see AERA, APA, & NCME, 1999; Kane, 2006). On the one hand, the use of think-aloud interviews in standardized achievement testing studies reflects a positive change in terms of broadening the types of data collected for generating strong validity arguments. On the other hand, with the exception of a few relevant studies by Norris (1988, 1990, 1991), there have not been experimental investigations of the accuracy of responses and consistency of processes in verbal reports when interviewer and item variables are manipulated in a performance-oriented, educational measurement context.

Standardized Achievement Test Items

There are at least two reasons to raise questions about the accuracy of responses and consistency of processes as measured by verbal reports in standardized achievement testing studies. First, standardized achievement test items, as a class of stimuli, could be expected to elicit unease or nervousness in examinees because these items are often associated with a single correct answer and used to evaluate student success in school classrooms. Although Ericsson and Simon (1993) reviewed a comprehensive list of psychological studies showing that the accuracy and consistency of response processes in verbal reports were not compromised when specific procedures were followed, most of the studies they reviewed involved psychological tasks and not standardized achievement test items.

The distinction between standardized achievement test items and psychological tasks should not be trivialized because it is well known that many students have relatively strong emotional reactions to standardized tests (Beilock & Carr, 2001; Beilock, Kulp, Holt, & Carr, 2004; Ericsson, 2006; Lewis & Linder, 1997; Ryan & Ryan, 2005). Unlike standardized achievement test items, psychological tasks are often viewed as "toy" tasks with relatively inconsequential effects for participants. Often unfamiliar, psychological tasks would not be expected to be associated with a single correct answer or used to evaluate participants in a classroom setting (see Ericsson & Simon, 1993). In other words, one would not expect psychological tasks to cause the same level of nervousness in students as standardized achievement test items.

Verbal reports collected in response to standardized achievement test items may be more vulnerable to inaccuracies, inconsistencies or both relative to verbal reports collected in response to psychological tasks. Working memory is the primary source of information for examinees as they think aloud (especially the concurrent verbal report, see Ericsson & Simon, 1993; also Leighton, 2004) and disruptions to working memory from nervousness could compromise the integrity of the reports (Wilson, 1994). It is well known that many students suffer from test anxiety or nervousness in performance-oriented situations (see Sawyer & Hollis-Sawyer, 2005). In some cases, students may worry excessively about the consequences of getting the wrong answer and this could interfere with their response processing and performance (Wolf & Smith, 1995). For example, students have been found to exhibit disruptions in response processing and suboptimal outcomes when solving standardized achievement test items because of excessive self-monitoring and exacerbating anxiety (see Beilock & Carr, 2001; Beilock et al., 2004; Ericsson, 2006; Lewis & Linder, 1997; Ryan & Ryan, 2005). Alternatively, students with strong metacognitive skills may know how to self-regulate their anxiety in performance-oriented situations (O'Neil & Abedi, 1996; Sundre & Kitsantas, 2004; Wolf, Smith, & Birnbaum, 1995). We discuss this possibility in the discussion section of the present article.

Interviewers

A second reason to question the accuracy of responses and consistency of processes as measured by verbal reports in standardized achievement testing studies is related to the presence of the interviewer. Although Ericsson and Simon (1993) recommended that the interviewer *not be* present during the interview, researchers often fail to follow this recommendation for fear that it will make the interview unnatural for examinees. Having an examinee think aloud in front of an interviewer is assumed to be more natural, akin to a teacher working one-on-one with a student, than having an examinee think aloud alone in a room and being watched from a video camera or two-way mirror. However, the naturalness that is sought from having the interviewer present during the interview may, in fact, increase the examinee's level of nervousness. For example, examinees could perceive the interviewer as a teacher who is focused on the correctness of the answers in order to make judgements about intelligence.

Although expert teachers, particularly in math, facilitate students' problem solving relative to less knowledgeable teachers (e.g., Darling-Hammond, 1996; Goldhaber & Brewer, 2000; Monk & King, 1994), the think-aloud interview may not offer sufficient time for examinees to develop the kind of trust that might exist with an expert teacher. Moreover, the think-aloud interview does not involve teaching students but rather listening, probing, and asking students questions about their problem solving. Students who perceive a performance-oriented situation as threatening to their self-worth may deflect attention away from their true ability and engage in alternative problem-solving behaviors that mask what they know or do not know (Butler & Neuman, 1995; Cain & Dweck, 1995; Covington, 1992; Pintrich, 2000; Ryan & Pintrich, 1997; Turner et al., 2002). Further, if students perceive the learning climate to be performance-oriented and lacking in trust, they are more likely to use self-handicapping tactics (e.g., Midgley & Urdan, 2001).

The presence of an interviewer could make students feel uneasy about the scrutiny that their performance will receive during the interview. For example, Norris (1990) found a statistically significant interviewer effect in a study of students' response processes to a critical thinking test. In this study, Norris had two interviewers conduct the interviews. With interviewer A, male and female students received identical thinking scores as measured by their verbal reports in response to the critical thinking test. However, with interviewer B, male students scored higher thinking scores than female students. Norris does not describe the characteristics of the interviewers. However, there is reason to hypothesize that an interviewer effect would also be found with standardized achievement test items as these items are performance-oriented and often used to make judgements about student success in school. Importantly, Norris (1990, p. 53) concluded from his study that "interviewer-related effects suggest one limitation to the verbal reporting procedure . . . this phenomenon is worthy of further study."

Three research questions are investigated in the present study: First, what are the effects of interviewer gender, interviewer knowledge level, and item difficulty on the *accuracy* of examinees' responses to test items in math? Second, what are the effects of these variables on examinees' nervousness? Third, what are the effects of interviewer gender, interviewer knowledge level, and item difficulty on the *consistency* of examinees' response processes in math as measured by their concurrent and retrospective verbal reports? In relation to these three questions, we also explored how students' gender and previous achievement in math influenced their response processing accuracy and consistency.

METHOD

Participants

Participants comprised a sample of convenience, which included 71 students (39 girls and 32 boys; mean age = 17.14 years, *SD* = .61 years) enrolled in a grade 12 university-tracked pure math course. Thirty-eight percent self-identified as Caucasian/White, 21% as Asian/Chinese, 16% as Filipino, 10% as South Asian or Southeast Asian, and the remaining 15% self identified as other ethnicities. Students were recruited from two academic high schools in a medium-sized metropolitan center. Students were required to have parental permission to participate and they were compensated for their time with a book gift certificate.

Materials

Individual think-aloud interviews were conducted with 71 students. Students were presented with a booklet that contained five sections in the following order:

Booklet section one. Section one contained 15 multiple-choice standardized achievement test items (each with a stem and four options). The 15 test items were sampled from a government-administered large-scale assessment in pure math. Items on this assessment were professionally developed and aligned with the curriculum of the grade 12 pure math course; the course in which participating students were enrolled. Two expert government-employed test developers, with over 15 years of experience assisted in selecting items that had acceptable levels of item discrimination and difficulty values that varied from easy to challenging (see Design). Although the items are secure and cannot be shown, the items included topics such as statistics, trigonometry and calculus. Each item was presented on a separate sheet, allowing students to use the empty space on the page as scratch paper if needed.

After students completed each item by circling an alternative on the page, students were asked to rate the item along two dimensions—familiarity and confidence. Students were asked "using the scale below where 0% means 'not at all familiar' and 100% means 'absolutely familiar,' please circle a value indicating how *familiar you are with the information* in this item." Students were also asked to do the same for *confidence in their solution* using a similar scale. The scales had 10 steps between 0 and 100%. Scales such as these have been used in previous studies to check whether students have had an opportunity to learn the material presented and are confident about their problem solving (see Cohen & Snowden, 2008).

Booklet sections two, three, four, and five. Section two included a 20-item state-based questionnaire of meta-cognition called the *Self-Assessment Questionnaire* (O'Neil & Abedi, 1996), where students responded to statements (e.g., I was aware of my own thinking) using a four-point Likert-type scale (i.e., not at all, somewhat, moderately so, and very much so). Section three included a 20-item state-based questionnaire of test anxiety called the *Test Attitude Inventory* (Spielberger et al., 1980), where students responded to statements (e.g., I feel confident and relaxed while taking tests) using a four-point Likert-type scale (i.e., almost never, sometimes, often, almost always). Both of these measures have internal consistency levels above .80 and have been validated for use with high-school seniors (see O'Neil & Abedi, 1996; Spielberger et al., 1980).

Section four included an 8-item questionnaire designed to measure students' perceptions of the think-aloud interview. Students responded using a seven-point Likert-type scale anchored by "strongly disagree" and "strongly agree." The critical item in this questionnaire probed students about whether they believed the interviewer was an expert in math (i.e., "I viewed the interviewer as an expert in mathematics"). This question was included as an attempt to check our manipulation of interviewer knowledge level (see Design). Finally, section five included a 4-item demographic questionnaire that requested information about students' last report card grade in math and English. The accuracy of students' reported grades was checked against school records. Booklet sections one, two, three, and four involved a selected-choice response format. Students were instructed to complete the sections of the booklet in the order presented and were not allowed to skip sections or "look ahead" to other sections.

Design

A fully crossed experimental design with two between-subject independent variables (i.e., interviewer gender and interviewer knowledge level) and one within-subject independent variable (i.e., item difficulty) was implemented as follows:

Overview of independent variables. Interviewer gender comprised two levels. Two graduate student assistants, one male and one female, were trained to conduct all 71 interviews. The two interviewers were Caucasian, approximately the same age (i.e., late twenties, early thirties), spoke English as their first language, conducted interviews in casual dress and were trained to establish and maintain a professional demeanour with students (e.g., no joking or informal chatting). Both assistants followed verbatim instructions for all think-aloud interviews (Ericsson & Simon, 1993).

Interviewer knowledge level comprised three levels—low knowledge (i.e., novice), high knowledge (i.e., expert), and control/neutral (i.e., neither novice nor expert). Interviewer knowledge level was manipulated by varying a portion of the verbatim instructions delivered to students at the start of the think-aloud interview. The full verbatim instructions are not reproduced here due to space limitations but the portion that varied across levels, labelled section B, is shown as follows for the *control/neutral* condition:

A. *Thank you for taking part in today's study . . . Now, before I explain what we will be doing today, let me introduce myself . . .*

B. ***My name is ________________ and I'm from the University of _________. And I'll be conducting the interview today.***

C. *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 . . .*

Instructions for the novice condition were identical to the control/neutral condition except the statements at section B read as follows:

> *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.*

Likewise, the instructions for the expert condition were identical to both other conditions except the statements at section B read as follows:

> *My name is __________ and I'm from the University of ________. My area of expertise is in Mathematics and I've been interested in how students solve problems for many years.*

Test item difficulty comprised three levels—easy, moderate, and difficult. As mentioned previously (see Materials), 15 test items were sampled to reflect easy, moderate, and difficult items as indicated by classical test theory *p*-values (i.e., proportion of students answering an item correctly). Five "easy" items were chosen to reflect p-values greater than 0.7; five "moderate" items were chosen to reflect *p*-values between .4 and .7; and five "difficult" items were chosen to reflect *p*-values less than .4.

Procedure. A sampling schedule was generated so that approximately equal numbers of participating girls and boys were randomly assigned to each of the six experimental conditions formed by crossing two between-subject variables (interviewer gender [2 levels] and interviewer knowledge [3 levels]). The two graduate assistants interviewed equivalent numbers of boys and girls. All students completed all levels of item difficulty, as this was a within-subject variable. The two graduate assistants began each individual interview by presenting a booklet to a student, self-identifying as the interviewer and explaining the objective of the interview. All interviews were conducted in a quiet room at the students' schools. Students were requested to think-aloud concurrently and retrospectively only for the 15 test items but not for the remaining four sections of the booklet. Standard concurrent interview probes ("Remember I'm just interested in your thoughts, please continue talking") and standard retrospective probes ("Please tell me all that you can remember about how you solved this problem") were used (Ericsson & Simon, 1993). Interviews lasted 45 minutes to 1 hour.

Overview of dependent variables. The present study focused on the following *quantitative* data: students' item responses to the 15 test items, which were scored dichotomously as 1 = correct and 0 = incorrect; familiarity scale ratings, confidence scale ratings, meta-cognitive scores, and test anxiety scores. The following *qualitative* data was also a focus of analysis: concurrent and retrospective reports from 71 students in response to each of 15 test items (2 types of reports × 15 items × 71 students = 2,130 verbal reports). Using guidelines outlined by Ericsson and Simon (1993), verbal reports were coded for frequency of nervous speech and level (sophistication) of response processing.

Nervous speech. Coding concurrent and retrospective reports independently for nervous speech was explored as a means to measure students' level of unease during the think-aloud

		Question: Concurrent	Question: Retrospective	Question: Concurrent	Question: Retrospective	Question: Concurrent	Question: Retrospective
Time		Start time: End time:	Start time: End time:	Start time: End time:	Start time: End time:	Start time: End time:	Start time: End time:
Aberrations in Speech Flow		▨	▨	▨	▨	▨	▨
a.	Word/sentence repetition						
b.	Ah/Er/Um/Hmm						
c.	Stutter						
d.	Fillers						
Validation Seeking Comments		▨	▨	▨	▨	▨	▨
e.	Validation seeking speech						
f.	Apologizing						
g.	Asking for clarifications						
h.	Delayed tactics						
Sentence Errors		▨	▨	▨	▨	▨	▨
i.	Sentence change						
j.	Sentence incompletion						
k.	Sentence correction						
l.	Speech Errors						
Paralanguage		▨	▨	▨	▨	▨	▨
m.	Goes into mumbling						
n.	Voice cracking						
o.	Throat clearing/Nervous laughter						
Interviewer Prompts		▨	▨	▨	▨	▨	▨
		▨	▨	▨	▨	▨	▨
p.	Prompts						

FIGURE 1 Checklist for rating occurrence of nervous speech in verbal reports.

interview. We searched the literature for a scale designed to measure nervous speech patterns. We did not find a scale for think-aloud interviews; therefore, as shown in Figure 1, we developed a checklist based on a review of clinical measures designed to evaluate anxiety in patients (Dibner, 1956; Krause & Pilisuk, 1961; Mahl, 1956, 1987). Clinical psychologists have found that disrupted patterns of speech are often indicative of nervousness and anxiety (Dibner, 1956; Krause & Pilisuk, 1961). Our review of clinical measures revealed the following four categories of speech that often indicate nervousness in patient populations:

1. *Aberrations in speech flow* such as word or sentence reiterations (e.g., use of *Ah* or *Er* or *Um* or *Hmm*), stuttering or repetition of words, and fillers such as extra words that do not convey meaning (e.g., "like" or "you know" or "well");
2. *Validation seeking comments* such as apologies, requests for clarification (e.g., "Pardon me?" "Is this the last question?"), and use of delay tactics (e.g., "This section is not my strong area," "I totally forgot how to do this");
3. *Sentence errors* such as sentence changes where a thought is started but then changes mid-way through completion of an idea (e.g., "I would like . . . I don't think I can solve this problem . . ."), sentence incompletion where an expression is interrupted without follow through (e.g., "And then I would . . ."), sentence self-corrections or grammatical corrections (e.g., "The reason I don't . . . didn't . . ."), and pronunciation errors; and
4. *Paralanguage* such as mumbling, voice cracking, and throat clearing (including laughter).

Two graduate student assistants, who were unaware of the experimental conditions and purpose of the study, were trained to independently count the frequency of nervous speech in students' concurrent and retrospective reports using Figure 1. The author trained the assistants using two randomly selected student interviews (each interview consisting of 15 concurrent reports and 15 retrospective reports for a total of 60 verbal reports). Considering the assistants as random or interchangeable "raters or judges," a two-way random effects calculation of intraclass correlation (Shrout & Fleiss, 1979) was used to evaluate the correlation between the assistants' counts. The correlation was .98 for concurrent reports and .95 for retrospective reports. Given the labour of coding concurrent and retrospective reports in response to each of 15 test items for 71 students, one assistant coded 35 interviews and the second assistant coded 36 interviews.

Cognitive models. Concurrent and retrospective reports were coded independently for level or sophistication of response processing. The two test developers who had assisted us in selecting items were asked to design cognitive models of *moderate*-ability and *high*-ability processing for responding correctly to each of the 15 test items. Previous research has shown that a single cognitive model reflecting one level of ability may not adequately describe the knowledge and skills of students at varying ability levels (see Leighton et al., 2009). High-ability students may use cognitive models that have fewer identifiable knowledge and skills than moderate-ability students because processes are "chunked" or automated (Leighton et al., 2009).

Four cognitive models were thus developed for each of the 15 test items (i.e., test developer A's moderate-ability model [TAM], test developer A's high-ability model [TAH], test developer B's moderate-ability model [TBM], and test developer B's high-ability model [TBH]). The cognitive models were not validated with an independent sample of students before being used in the present study because the purpose of the study was in part to determine whether these cognitive models captured differences in response processing for our sample of students. As an example, shown in Figure 2 are cognitive models of moderate ability and high ability for solving one of the easy items used in the study. Overall, 60 cognitive models were created (2 levels of cognitive models × 15 test items × 2 test developers).

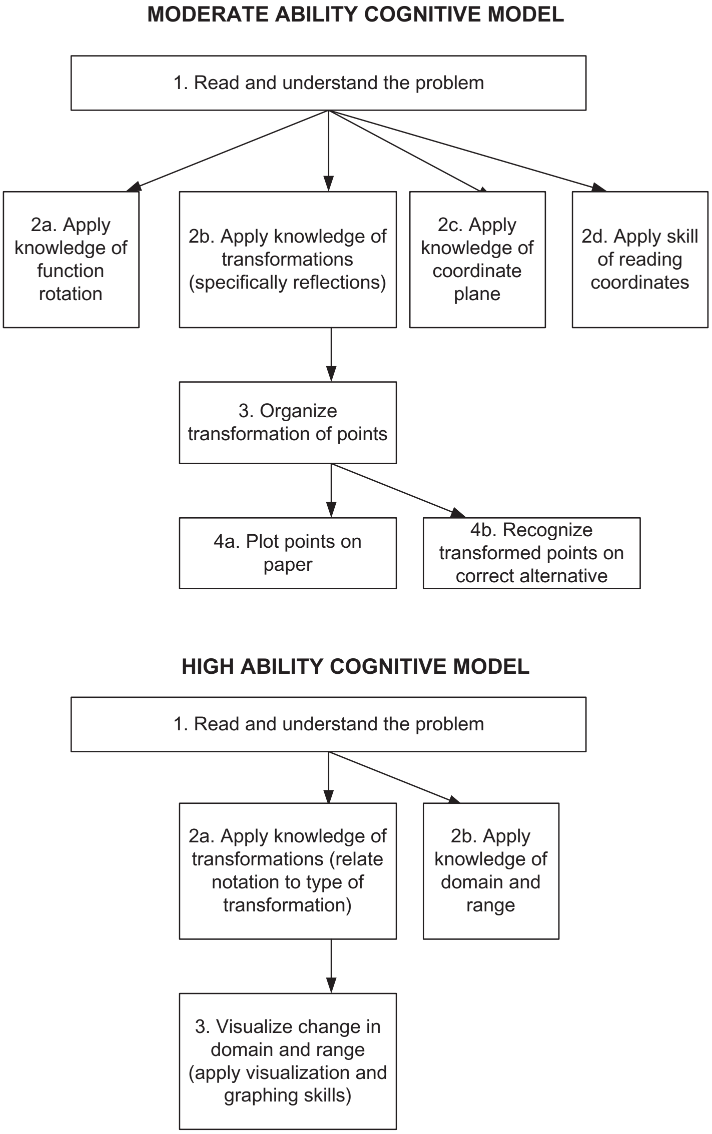
MODERATE ABILITY COGNITIVE MODEL
1. Read and understand the problem
2a. Apply knowledge of function rotation
2b. Apply knowledge of transformations (specifically reflections)
2c. Apply knowledge of coordinate plane
2d. Apply skill of reading coordinates
3. Organize transformation of points
4a. Plot points on paper
4b. Recognize transformed points on correct alternative
HIGH ABILITY COGNITIVE MODEL
1. Read and understand the problem
2a. Apply knowledge of transformations (relate notation to type of transformation)
2b. Apply knowledge of domain and range
3. Visualize change in domain and range (apply visualization and graphing skills)

FIGURE 2 Moderate-ability (top) and high-ability (bottom) models for easy item developed by test developer.

Two graduate student assistants (different from those who coded nervous speech and also unaware of the experimental conditions) were trained to identify response processes in concurrent and retrospective reports. Assistants compared response processes in concurrent and retrospective reports independently to the response processes outlined in the cognitive models designed by the test developers. The assistants then classified the reports (again independently) into a model category (i.e., TAM, TAH, TBM, TBH) that reflected a majority of matching response processes. The four cognitive models were expected to reflect response-processing differences in the 71 students' concurrent and retrospective reports. If none of the four models described the response processes in a student's verbal report, the report was classified as "no model." Assistants were trained using three student interviews (2 types of reports × 15 test items × 3 students = 90 verbal reports) and their inter-rater agreement was calculated. Because classifying concurrent and retrospective reports into cognitive model categories (including no model) required the assistants to make mutually exclusive classifications, Cohen's kappa was calculated. A kappa value of .81 was obtained for assistants' inter-rater agreement, and a value of 1.0 was obtained after disagreements were discussed. Most of the disagreements were minor and included misinterpreting a part of a student's verbal report. After training on the three interviews, both assistants listened to all 71 interviews and classified all 2,130 reports (2 types of reports × 15 test items × 71 students) in order to minimize any misclassification.

RESULTS

In this section we describe the results of manipulating interviewer gender, interviewer knowledge level, and item difficulty on students' (a) item response accuracy, familiarity and confidence ratings, meta-cognition and test anxiety scores, (b) nervous speech during concurrent and retrospective reports, and (c) consistency of cognitive models (response processes) used in concurrent and retrospective reports. We do not include student gender as a predictive variable in the following presentation because preliminary analysis revealed that it did not have an effect on any of the dependent variables, including item response accuracy, meta-cognition, test anxiety, nervous speech, or cognitive models. Further, student gender did not interact with interviewer gender, interviewer knowledge level or item difficulty.

Accuracy, Familiarity, Confidence of Responses, Meta-Cognition, and Test Anxiety

To evaluate the effects of interviewer gender, interviewer knowledge level, and item difficulty on students' item response accuracy, a 2 (interviewer gender) × 3 (interviewer knowledge level) × 3 (item difficulty) mixed ANOVA with repeated measures on the last factor was conducted. Table 1 shows the means and standard deviations for item response accuracy by interviewer gender, interviewer knowledge level, and item difficulty. The ANOVA results indicated that there was no significant main effect of interviewer *gender* on students' item response accuracy. However, there was a statistically significant main effect of interviewer *knowledge level,* $F(2, 65) = 4.57$, $p < .05$, $\omega^2 = .12$, with Scheffe post-hoc tests revealing that students assigned to an interviewer who identified himself or herself as an *expert* in math performed *less well* than students assigned to an interviewer who identified himself or herself as a novice (mean difference of .63, SE = .21, $p < .05$). There was no statistical difference in students' item response accuracy between the expert and control condition, and the control and novice condition. However, students in the control condition performed less well than students in the novice condition but better than students in the expert condition for moderate and difficult items. Further, in response to the question "I viewed the interviewer as an expert in mathematics" in section four of the booklet, students rated

TABLE 1
Means and Standard Deviations for Students' Accuracy, Familiarity Ratings, and Confidence Ratings on Easy, Moderate, and Difficult Items by Interviewer Gender and Interviewer Knowledge Conditions

	Expert		*Control/Neutral*		*Novice/Nonexpert*	
	M	*F*	*M*	*F*	*M*	*F*
Easy	2.73 (1.00)	3.50 (1.08)	4.18 (.98)	3.69 (.94)	3.75 (.86)	3.75 (.96)
Mod	2.27 (1.27)	2.92 (.79)	2.55 (1.21)	3.38 (1.50)	3.33 (.98)	3.25 (1.13)
Hard	1.82 (.75)	1.83 (.83)	1.91 (1.13)	1.62 (1.12)	2.50 (1.38)	2.33 (1.15)
F-Easy	68.45 (15.50)	70.50 (16.20)	77.97 (12.96)	73.23 (18.30)	74.33 (17.99)	78.08 (13.59)
F-Mod	66.27 (14.53)	68.00 (22.99)	76.00 (10.27)	75.85 (15.82)	74.33 (14.86)	79.17 (16.70)
F-Hard	61.00 (16.88)	71.50 (22.37)	73.68 (14.80)	72.00 (16.24)	71.16 (17.79)	75.08 (16.00)
C-Easy	62.36 (11.36)	69.50 (17.27)	68.69 (14.07)	67.07 (15.13)	64.00 (19.96)	72.66 (9.40)
C-Mod	62.63 (16.24)	60.33 (22.83)	66.36 (14.52)	64.82 (18.79)	65.00 (19.86)	67.91 (16.12)
C-Hard	58.72 (13.80)	68.83 (17.73)	63.31 (16.29)	59.61 (17.73)	66.50 (21.76)	65.91 (16.96)

Notes. M = Male Interviewer, F = Female Interviewer; F-Easy = Familiarity ratings for easy items, F-Mod = Familiarity ratings for moderate items, F-Hard = Familiarity ratings for difficult items; C-Easy = Confidence ratings for easy items, C-Mod = Confidence ratings for moderate items, and C-Hard = Confidence ratings for difficult items; Accuracy scores range from 0 to 5, familiarity ratings range from 0 to 100, and confidence ratings range from 0 to 100. Sample sizes within the six experimental conditions ranged from 11–12 students.

the interviewer as an "expert" more often in the control and expert conditions than in the novice condition, $t(1, 69) = 2.629$, $p < .01$).

The ANOVA results also indicated a statistically significant main effect of item difficulty, $F(2, 130) = 47.46$, $p < .001$, $\omega^2 = .42$; not surprisingly, students' item response accuracy was better for easy items than for moderate items, and students' item response accuracy was better for moderate items than for difficult items (tests of within-subject contrasts, $F[1, 65] = 12.53$, $p < .001$, $\omega^2 = .16$; $F[1,65] = 38.67$, $p < .001$, $\omega^2 = .37$, respectively). There were no other statistically significant effects, including interactions between variables. Although Cronbach's alpha for the 15 items was calculated at .43, it is not clear how useful this value is given that interviewer knowledge had an effect on students' item response accuracy so the items were no longer simply measuring a single dimension.

A 2 (interviewer gender) × 3 (interviewer knowledge level) × 3 (item difficulty) mixed ANOVA with repeated measures on the last factor was also conducted on familiarity and confidence ratings. There were no statistically significant effects of interviewer gender, interviewer knowledge level, or item difficulty on either familiarity or confidence ratings. In other words, students expressed approximately the same familiarity with the content of all 15 items irrespective of the experimental condition to which they were assigned (see Table 1). Likewise, students expressed approximately the same confidence in their solutions for all 15 items irrespective of the experimental condition to which they were assigned (see Table 1). Table 2 shows the correlations among item response accuracy, item familiarity and item confidence ratings across levels of item difficulty. As expected these variables shared moderate to strong associations. Cronbach's alphas for familiarity and confidence ratings were, respectively, .86 and .82.

A similar mixed ANOVA was also conducted on meta-cognition and test anxiety scores and there were no effects of interviewer gender, interviewer knowledge level, or item difficulty on

TABLE 2
First Order Correlations Among Item Response Accuracy and Item Familiarity and Item Confidence Ratings

Measure	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*M*	*SD*
1. E. Accuracy	—	.41**	.43**	.06	.19	.09	.19	.22	.15	3.58	1.04
2. E. Familiarity	.41**	—	.79**	−.01	.59**	.22	.24*	.51**	.42**	74.43	15.51
3. E. Confidence	.43**	.79**	—	−.01	.45**	.40**	.33**	.42**	.49**	67.94	14.68
4. M. Accuracy	.06	−.01	−.01	—	.25*	.43**	.39**	.24*	.35**	3.00	1.20
5. M. Familiarity	.19	.59**	.45*	.25*	—	.72**	.32**	.84**	.66**	73.75	16.53
6. M. Confidence	.09	.22	.40**	.43**	.72*	—	.47**	.57**	.69**	64.65	18.11
7. D. Accuracy	.19	.24*	.33**	.39**	.32**	.47**	—	.34**	.51**	1.99	1.11
8. D. Familiarity	.22	.51**	.42**	.24*	.84**	.57**	.34**	—	.75**	71.57	17.20
9. D. Confidence	.15	.42**	.49**	.35**	.66**	.69**	.51**	.75**	—	64.50	17.15
M	3.58	74.43	67.94	3.00	73.75	64.65	1.99	71.57	64.50		
SD	1.04	15.51	14.68	1.20	16.53	18.11	1.11	17.20	17.15		

Notes. E = easy items, M = moderately difficult items, and D = difficult items. $N = 71$. $^*p < .05$, $^{**}p < .01$.

TABLE 3
Means and Standard Errors for Students' Metacognition and Test Anxiety by Interviewer Knowledge and Interviewer Gender Conditions

	Expert		*Control/Neutral*		*Novice/Nonexpert*	
	M	*F*	*M*	*F*	*M*	*F*
SA	55.37 (3.26)	59.41 (2.66)	64.20 (2.91)	59.30 (2.55)	58.40 (2.91)	61.91 (2.66)
TAI	39.37 (4.45)	44.83 (3.63)	39.80 (3.98)	38.92 (3.49)	41.20 (3.98)	44.58 (3.63)

Notes. M = Male Interviewer, F = Female Interviewer; SA = Self-Assessment Questionnaire (scores range from 20 to 80), TAI = Test Attitude Inventory (scores range from 20 to 80). Sample sizes within the six experimental conditions ranged from 11–12 students.

students' scores. Table 3 shows their means and standard deviations. Cronbach's alphas for metacognition scores and test-anxiety scores were, respectively, .87 and .92. It is important note that although there were no effects of interviewer gender and knowledge level on test anxiety scores, this does not necessarily indicate that the think aloud interview was not a high-stakes or performance-oriented situation for students. There were significant interviewer effects on item response accuracy (students performed more poorly in the expert condition). Thus, it is possible that the moderate-ability and high-ability students in our sample were able to self-regulate their emotions in this performance-oriented situation despite a reduction in response accuracy for students in the expert condition (see Discussion).

Given these initial results, further analyses were subject to the following considerations. First, because interviewer gender did not have an effect on students' item response accuracy, familiarity, confidence, or meta-cognitive and test anxiety scores, it was not included in further analyses. Second, given that students' item response accuracy in the expert condition did not differ significantly from the control/neutral condition, and students in both these conditions perceived the interviewer to be a math expert, the conditions were combined. Although the sample size in the expert plus control condition ($n = 47$) and the novice condition ($n = 24$) differed in number, within-group variances between the conditions remained statistically equivalent. Third, in order to evaluate the effect of students' previous achievement in pure math on their item response accuracy, we formed a binary variable by dividing students into two groups. The groups were moderate achieving students ($n = 35$) with a last report card grade of less than or equal to 81% in pure math and high achieving students ($n = 36$) with a last report card grade greater than or equal to 82% in pure math. We did not include students' English scores because not all students were in the same grade level of English classes.

Next, we conducted a 2 (interviewer knowledge level—expert/control versus novice) × 2 (previous student achievement—moderate high versus high) × 3 (item difficulty) mixed ANOVA with repeated measures on the last factor on students' item response accuracy. As shown in Table 4, interviewer knowledge continued to have a statistically significant effect on students' item response accuracy, $F(1, 67) = 5.92$, $p < .05$, $\omega^2 = .08$; that is, students assigned to the expert/control conditions performed *less well* than students assigned to the novice condition. Students' previous achievement also had an effect on item accuracy, $F(1, 67) = 14.21$, $p < .001$, $\omega^2 = .17$; not surprisingly, students who were moderate achieving in math performed less well than students who were high achieving. Finally, there was an effect of item difficulty on students' item response accuracy, $F(2, 134) = 39.14$, $p < .001$, $\omega^2 = .36$; as expected, easy items were significantly easier than moderate items, and moderate items were significantly easier than difficult items (tests of within-subject contrasts, $F[1, 67] = 8.82$, $p < .01$, $\omega^2 = .11$; $F[1,67] = 35.08$, $p < .001$, $\omega^2 = .34$, respectively). No other effects were found.

Nervous Speech

The frequency of nervous speech was evaluated in concurrent reports and retrospective reports using the checklist shown in Figure 1. The frequency distributions for aberrations in speech flow, validation-seeking comments, sentence errors, and paralanguage were positively skewed. To normalize the distributions, a natural logarithmic transformation was applied to the data to stretch out the lower ends of the distributions and compress the upper ends. Once this was done, a 2 (interviewer knowledge level) × 2 (previous student achievement) × 3 (item difficulty) mixed ANOVA with repeated measures on the last factor was conducted. Only two significant effects

TABLE 4
Means and Standard Deviations for Students' Response Accuracy on Easy, Moderate, and Difficult Items by Interviewer Knowledge and Previous Achievement in Pure Mathematics

	Expert + Control		*Novice/Nonexpert*	
	MH	*H*	*MH*	*H*
Easy	3.17 (1.20)	3.91 (.84)	3.55 (.93)	3.92 (.86)
Mod	2.38 (1.20)	3.26 (1.17)	3.09 (.94)	3.46 (1.12)
Hard	1.62 (.82)	1.96 (1.06)	1.82 (.87)	2.92 (1.32)

Notes. MH = Moderately-high achieving, H = High achieving; score range from 0 to 5. Sample sizes within the four experimental conditions ranged from 12–23 students.

TABLE 5
Means and Standard Deviations for Students' Transformed Scores on Aberrations in Speech Flow in Concurrent Reports on Easy, Moderate, and Difficult Items by Interviewer Knowledge and Previous Achievement in Pure Mathematics

	Expert + Control		*Novice/Nonexpert*	
	MH	*H*	*MH*	*H*
Easy	3.15 (1.00)	3.15 (.79)	3.35 (.57)	3.04 (.59)
Mod	3.09 (.77)	3.13 (.80)	3.10 (.94)	2.97 (.86)
Hard	2.58 (1.11)	3.01 (.59)	2.94 (.67)	2.80 (.78)

Notes. MH = Moderately-high achieving, H = High achieving; score range from 0 to 5. Sample sizes within the four experimental conditions ranged from 12–23 students.

TABLE 6
Means and Standard Deviations for Students' Transformed Scores on Validation Seeking Speech in Concurrent Reports on Easy, Moderate, and Difficult Items by Interviewer Knowledge and Previous Achievement in Pure Mathematics

	Expert + Control		*Novice/Nonexpert*	
	MH	*H*	*MH*	*H*
Easy	1.72 (.80)	1.85 (.79)	1.88 (.71)	1.41 (.63)
Mod	1.39 (.73)	1.66 (.76)	1.29 (.84)	1.15 (.71)
Hard	1.25 (.84)	1.46 (.88)	1.37 (.78)	.84 (.49)

Notes. MH = Moderately-high achieving, H = High achieving; score range from 0 to 4. Sample sizes within the four experimental conditions ranged from 12–23 students.

were found and both were found only for concurrent reports. First, item difficulty was found to have a significant effect on *aberrations in speech flow,* $F(2, 130) = 9.17$, $p < .001$, $\omega^2 = .12$; namely, students exhibited more aberrations in speech flow on easier and moderate items than on difficult items. Second, item difficulty was found to have a significant effect on *validation seeking comments,* $F(2, 132) = 15.61$, $p < .001$, $\omega^2 = .19$; namely, students sought more validation seeking comments on easier items than on moderate and difficult items. Shown in Tables 5 and 6 are the means and standard deviations for aberrations in speech flow and validation seeking comments. Cronbach's alpha was .81 across all four categories of nervous speech.

Cognitive Models

A full analysis of students' cognitive models is elaborated in Wang and Leighton (2011). For the purpose of this article, however, we only report the effects of interviewer knowledge, previous achievement in pure math, and item difficulty on students' cognitive models.

Students' concurrent and retrospective reports were independently classified into one of five cognitive model categories (including no model use). Concurrent and retrospective reports were then assigned numerical scores to index level or *sophistication of response processing*. For example, concurrent and retrospective reports reflecting *moderate-ability* cognitive models, irrespective of whether they matched test developer A's or B's models, were assigned a score of 1. Concurrent and retrospective verbal reports reflecting *high-ability* cognitive models, irrespective of whether they matched test developer A's or B's models, were assigned a score of 2. Concurrent and retrospective verbal reports reflecting no cognitive models were assigned a score of 0.

Next, aggregate cognitive model scores for students' concurrent and retrospective reports were compiled by item difficulty. For example, by summing all five cognitive model scores for concurrent reports to five easy items, an aggregate *concurrent cognitive model score* was created for easy items. We also created aggregate cognitive model scores for moderate items and difficult items. It is important to note that a student could have solved an item correctly and reveal no identifiable cognitive model; in this case the student's concurrent report or retrospective report or both were assigned a value of zero. These occurrences were rare as most classifications of "no model" occurred because the student did not know the answer or guessed. Although concurrent and retrospective cognitive model scores were positively correlated with response accuracy across item difficulty (statistically significant correlations from .24 to .44), our analysis focused on the level of response processing and consistency of cognitive model scores across concurrent and retrospective reports. We did not focus on the correctness of cognitive model applications since item response accuracy already reflected proper application to some degree. Further, it was important for us to distinguish between item response accuracy and cognitive models because it was possible for students to exhibit proper knowledge and skills in their response processing but still get the answer wrong because of computational mistakes. Cronbach's alpha for concurrent and retrospective cognitive model scores across items was .77.

A 2 (interviewer knowledge level) × 2 (previous student achievement) × 3 (item difficulty) mixed ANOVA with repeated measures on the last factor was conducted on concurrent and retrospective cognitive model scores. For *concurrent reports*, two statistically significant effects were found. First, previous student achievement had an effect on cognitive model scores exhibited in the reports, $F(1,67) = 22.37, p < .001, \omega^2 = .25$; namely, high achieving students exhibited more sophisticated response processing than moderate achieving students. Second, item difficulty had an effect on concurrent cognitive model scores, $F(2, 134) = 6.16, p < .01, \omega^2 = .084$; items of greater difficulty elicited more sophisticated response processing than items of lesser difficulty. There was no effect of interviewer knowledge level.

For *retrospective reports*, two statistically significant effects were found. First, higher achieving students exhibited higher cognitive model scores than moderate achieving students, $F(1, 67) = 23.21, p < .001, \omega^2 = .25$. Second, interviewer knowledge level interacted with item difficulty, $F(2, 134) = 3.43, p < .05, \omega^2 = .049$. As shown in Figure 3 and Table 7, students in the novice condition earned higher retrospective cognitive model scores than students in the expert/control condition but only for moderate and difficult items. For easy items, students in the expert/control condition earned higher retrospective cognitive model scores than students in the novice condition. Means and descriptive statistics for cognitive model scores displayed in concurrent and retrospective reports are shown in Tables 7 and 8.

Finally, the consistency of cognitive model scores across concurrent and retrospective reports was evaluated. A 2 (interviewer knowledge level) by 2 (previous student achievement) by 2 (verbal report type) mixed ANOVA with repeated measures on the last factor was conducted for each level of item difficulty. For easy items, students' concurrent and retrospective reports were found to be significantly different in the sophistication of response processing, $F(1, 67) = 7.48$,

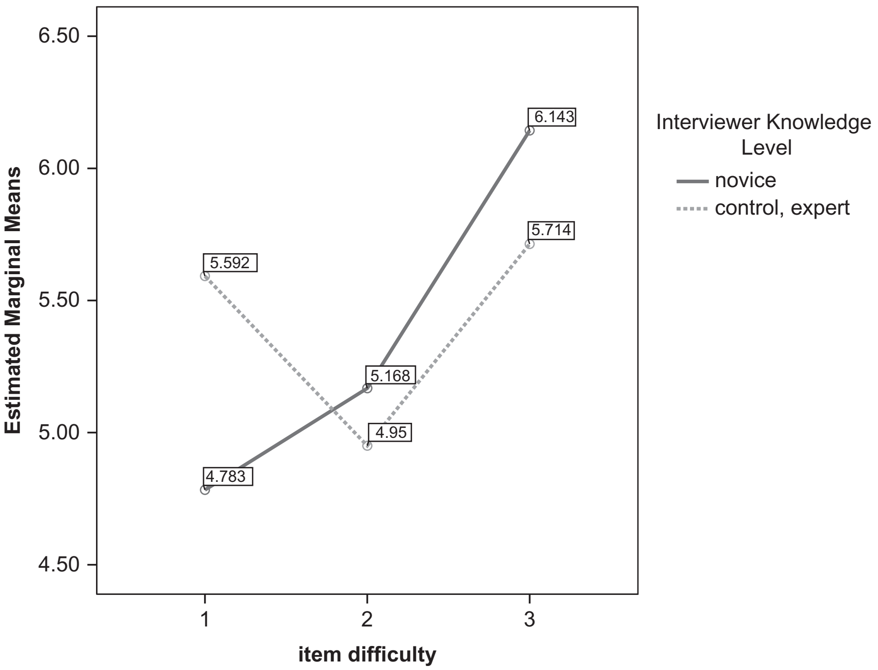

FIGURE 3 Interaction of interviewer knowledge level (Expert/Control versus Novice) and item difficulty (1 = easy items, 2 = moderate items, and 3 = difficult items) for cognitive model scores in retrospective reports (see Table 8).

TABLE 7
Means and Standard Deviations for Students' Cognitive Model Scores in Retrospective Reports by Item Difficulty, Interviewer Knowledge and Previous Achievement in Pure Mathematics

	Expert + Control		*Novice/Nonexpert*	
	MH	*H*	*MH*	*H*
Easy	4.75 (1.45)	6.43 (1.59)	4.18 (1.53)	5.38 (1.44)
Mod	4.29 (1.85)	5.60 (1.07)	5.18 (1.40)	5.15 (1.34)
Hard	5.16 (1.68)	6.26 (1.21)	5.36 (1.62)	6.92 (1.44)

Notes. MH = Moderately-high achieving, H = High achieving; score range from 0 to 10. Sample sizes within the four experimental conditions ranged from 12–23 students.

$p < .01$, $\omega^2 = .10$; as shown in Table 9, cognitive model scores were higher for retrospective reports than for concurrent reports. Also, for easy items, higher achieving students displayed more sophisticated response processing than moderate achieving students, $F(1, 67) = 18.09$, $p < .001$, $\omega^2 = .21$.

TABLE 8
Means and Standard Deviations for Students' Cognitive Model Scores in Concurrent Reports by Item Difficulty, Interviewer Knowledge and Previous Achievement in Pure Mathematics

	Expert + Control		*Novice/Nonexpert*	
	MH	*H*	*MH*	*H*
Easy	3.83 (1.71)	5.69 (1.57)	4.00 (1.94)	4.76 (1.16)
Mod	4.45 (1.84)	5.60 (1.58)	4.36 (2.76)	5.69 (1.25)
Hard	4.66 (1.80)	6.78 (1.38)	4.18 (2.60)	6.23 (1.36)

Notes. MH = Moderately-high achieving, H = High achieving; score range from 0 to 10. Sample sizes within the four experimental conditions ranged from 12–23 students.

TABLE 9
Means and Standard Deviations for Students' Cognitive Model Scores in Concurrent and Retrospective Reports by Item Difficulty and Previous Achievement in Mathematics

	Concurrent		*Retrospective*	
	MH	*H*	*MH*	*H*
Easy	3.88 (1.76)	5.36 (1.49)	4.57 (1.48)	6.05 (1.60)
Mod	4.42 (2.13)	5.63 (1.45)	4.57 (1.75)	5.44 (1.18)
Hard	4.51 (2.06)	6.58 (1.38)	5.22 (1.64)	6.50 (1.32)

Notes. MH = Moderately-high achieving, H = High achieving; score range from 0 to 10. Sample sizes within the four experimental conditions ranged from 12–23 students.

For moderate items, students' concurrent reports were found to be consistent with retrospective reports in response processing. Again, higher achieving students displayed higher cognitive model scores than moderate achieving students, $F(1, 67) = 7.08$, $p < .01$, $\omega^2 = .09$. For difficult items, students' concurrent and retrospective reports were found to be significantly different in response processing, $F(1, 67) = 4.49$, $p < .05$, $\omega^2 = .06$; cognitive model scores were again higher in retrospective reports than in concurrent reports. Also, higher achieving students displayed higher cognitive model scores than moderate achieving students, $F(1, 67) = 24.11$, $p < .001$, $\omega^2 = .26$. There were no interactions or other statistically significant effects.

DISCUSSION

The objective of the current study was to examine the accuracy and consistency of data obtained from think-aloud interviews in an educational measurement context. Our research questions were (a) what are the effects of interviewer gender, interviewer knowledge level, and item difficulty on the accuracy of examinees' response processes as measured by their item performance in pure math? (b) what are the effects of these variables on examinees' level of nervousness? and (c) what are the effects of interviewer gender, interviewer knowledge level, and item difficulty on the consistency of examinees' response processes in concurrent and retrospective verbal reports of these math items? We also explored how students' gender and previous achievement in pure math influenced their response processing accuracy and consistency.

One of the noteworthy findings in this study was that students who believed a math expert was interviewing them were less accurate in their responses than those students who believed a novice was interviewing them. However, the underlying mechanism of this effect is not clear. Although we hypothesized that an expert interviewer would make students nervous and disrupt their response processing, we did not find an effect of interviewer knowledge on students' nervous speech, test anxiety, or cognitive model scores in concurrent reports. We did find greater frequency of some types of nervous speech on easier items but given that verbal utterances are often more frequent for easier than more difficult items this result was not unexpected. One explanation is that all students regulated their nervousness—even those students in the expert/control condition who were performing less well than students in the novice condition. High achieving students have been found to successfully regulate their emotions in academic situations (Boekaerts & Corno, 2005). Our study included moderate to high achieving students and they reported moderate levels of meta-cognition across all interviewer conditions (SA total scores in Table 2).

Another explanation is that the disruptive effect of believing an expert is conducting the interview falls on *lower-level* response processing such as computational execution and not on *higher-level* response processing such as knowledge and skills. Recall that cognitive models were developed to show the response processing—knowledge and skills—required to solve items correctly. Higher-ability cognitive models reflected more sophisticated response processes than moderate-ability models but all models, if implemented properly, were expected to lead to correct answers. Students in the expert/control condition suffered in their item response accuracy but their level of response processing as indexed by concurrent cognitive model scores did not differ from students in the novice condition. Thus, it is possible that students in the expert condition did not implement their cognitive models properly. In other words, although students in the expert/control condition reported similar response processes as students in the novice condition, students in the expert condition may have experienced disruptions in execution of knowledge and skills; leading to basic computational mistakes and selecting the wrong answer. Future research could explore specifically the computational response processes that are disrupted or left intact by interview conditions.

Disruptions in computational execution of knowledge and skills may have also influenced students' retrospective reports. We found that students who believed an expert was interviewing them recalled *less* sophisticated response processing—knowledge and skills—in their retrospective reports for moderate and difficult items than students in the novice condition (see Figure 3). Ericsson and Simon (1993) indicate that retrospective reports are prone to bias because students have to recall how they solved the task. As discussed previously, it is possible that students, believing they were being interviewed by an expert in a performance-oriented situation, perceived a threat to their self-worth and deflected attention away from how they actually solved the problem to some alternative problem-solving behavior that masked what they knew or did not know (e.g., Cain & Dweck, 1995; Pintrich, 2000; Turner et al., 2002). Again, future studies need to explore the social desirability of students' verbal reports given assumptions made about the interviewer in performance-oriented situations.

The second main finding was the effect of item difficulty on the consistency of response processing in concurrent and retrospective reports. For easy and difficult items, concurrent and retrospective reports differed significantly in cognitive model scores, with students' recalling more sophisticated response processes in their retrospective reports than in concurrent reports. However, for moderate items, concurrent and retrospective reports were consistent in cognitive model scores. This latter result is in line with Ericsson and Simon's (1993) recommendation that the most informative verbal reports are collected from items of moderate difficulty.

The results from this study should not be taken to indicate that think aloud interviews ought not to be conducted in educational measurement studies. However, caution must be taken with this source of data. In particular, the characteristics of interviewers need to be considered when collecting verbal reports of standardized achievement test items. In this study, we found evidence that when interviewers present themselves as experts in the content domain of the interview or fail to say anything about their expertise, students' may assume they are experts and item response accuracy can deteriorate. Item response accuracy may deteriorate in part because processes—knowledge and skills—are not executed properly. Another area of caution is the effect of interviewer knowledge level on students' retrospective reports. Students who thought an expert was interviewing them recalled less sophisticated response processing in their retrospective reports for moderate and difficult items than students who thought a novice was interviewing them.

With think-aloud interviews used in educational measurement studies to build validity arguments for test-based inferences, care must be taken that the data obtained from verbal reports support accurate inferences about students' response processes and achievement. Research into the methodological aspects of think-aloud interviews is warranted if we are to interpret these data with the confidence required to judge examinees' response processes.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Institutes for Research (AIR). (2009). *Cognitive lab testing*. Retrieved from http://www.air.org/topics/topic_cognitive_lab_testing.aspx

Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, *130*, 701–725.

Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General*, *133*, 584–600.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An International Review*, *54*, 199–231.

Butler, R., & Neuman, O. (1995). Effects of task and ego achievement goals on help-seeking behaviors and attitudes. *Journal of Educational Psychology*, *87*, 261–271.

Cain, K. M., & Dweck, C. S. (1995). The relation between motivational patterns and achievement cognitions through the elementary school years. *Merrill-Palmer Quarterly*, *41*, 25–52.

Cohen, D. J., & Snowden, J. L. (2008). The relations between document familiarity, frequency, and prevalence and document literacy performance among adult readers. *Reading Research Quarterly*, *43*(1), 9–26.

Covington, M. V. (1992). *Making the grade*. Cambridge, England: Cambridge University Press.

Darling-Hammond, L. (1996). What matters most: A competent teacher for every child. Phi Delta Kappan, *77*, 193–201.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, *26*, 1–22.

Dibner, A. S. (1956). Cue-counting: A measure of anxiety in interviews. *Journal of Consulting Psychology*, *20*, 475–478.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert review. *Educational Measurement: Issues and Practice*, *29*, 24–35.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223–241). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–621). Westport, CT: National Council on Measurement in Education and American Council on Education.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*, 129–145.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: National Council on Measurement in Education and American Council on Education.

Krause, M. S., & Pilisuk, M. (1961). Anxiety in verbal behavior: A validation study. *Journal of Consulting Psychology*, *25*, 414–419.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*, 6–15.

Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchical consistency index. *Applied Measurement in Education*, *22*, 229–254.

Lewis, B., & Linder, D. (1997). Thinking about choking? Attentional processes and paradoxical performance. *Personality and Social Psychology Bulletin*, *23*, 937–944.

Mahl, G. F. (1956). Disturbances and silences in the patient's speech in psychotherapy. *Journal of Applied Social Psychology*, *53*, 1–15.

Mahl, G. F. (1987). Everyday disturbances of speech. In R. L. Russell (Ed.), *Language in psychotherapy: Strategies of discovery* (pp. 213–269). New York, NY: Plenum.

Midgley, C., & Urdan, T. (2001). Academic self-handicapping and performance goals: A further examination. *Contemporary Educational Psychology*, *26*, 61–75.

Monk, D. H., & King, J. (1994). Multi-level teacher resource effects on pupil performance in secondary mathematics and science: The role of teacher subject matter preparation. In R. G. Ehrenberg (Ed.), *Contemporary policy issues: Choices and consequences in education* (pp. 29–58). Ithaca, NY: ILR Press.

Norris, S. P. (1988). Controlling for background beliefs when developing multiple-choice critical thinking tests. *Educational Measurement*, *7*, 5–11.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking performance. *Journal of Educational Measurement*, *27*, 41–58.

Norris, S. P. (1991). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 451–472). Hillsdale, NJ: Erlbaum.

O'Neil, H. F., & Abedi, J. (December, 1996). *Reliability and Validity of a State Metacognitive Inventory: Potential for Alternative Assessment*. CSE Technical Report 469. National Center for Research on Evaluation Standards and Student Testing (CRESST). University of California, Los Angeles, CA.

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*, 544–555.

Ryan, A. M., & Pintrich, P. R. (1997). Should I ask for help? The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology*, *89*, 329–341.

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist*, *40*, 53–63.

Sawyer T. P., Jr., & Hollis-Sawyer, L. A. (2005). Predicting stereotype threat, test anxiety, and cognitive ability test performance: An examination of three models. *International Journal of Testing*, *5*, 225–246.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *2*, 420–428.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York, NY: American Council on Education, Macmillan.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, E. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Manual for the test anxiety inventory* ("Test Attitude Inventory"). Redwood City, CA: Consulting Psychologists Press.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6–26.

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics. *Journal of Educational Psychology*, *94*, 88–106.

Wang, X., & Leighton, J. P. (2011). *Evaluating four coding schemes for categorizing cognitive models in verbal reports for educational measurement studies*. Unpublished manuscript.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science*, *5*, 249–252.

Wolf, L. F. & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227–242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation and mentally taxing items. *Applied Measurement in Education*, *8*, 341–352.

Zucker, S., Sassman, C., & Case, B. J. (February, 2004). *Cognitive Labs*. Technical Report. Pearson Education. Retrieved from http://pearsonassess.com/NR/rdonlyres/E5CD33E6-D234-46F3-885A-9358575372FB/0/CognitiveLabs_Final.pdf