A peer reviewed, open-access electronic journal: ISSN 1531-7714

Considerations for Designing Measures of Confidence

Casandra Koevoets-Beach, *University of Louisville*
Donya Kurdi, *University of Louisville*
Morgan Balabanoff, *University of Louisville*

Abstract: Confidence tiers have been paired with multiple choice items across different fields since the early twentieth century and have seen widespread adoption in discipline-based education research fields seeking to evaluate aspects of self-regulated learning. The design of two-tiered confidence judgments impacts interpretability and perception of their utility, so meaningful engagement hinges on intentional design for specific constructs. This study uses cognitive interviews exploring students' interpretations of confidence tier components to identify design features which elicit meaningful variation in self-reflection. The evidence supports two prevailing motivations for using confidence tiers—prompting metacognition and measuring the strength of learners' alternate conceptions. The challenges and strategies students encounter while engaging with confidence tiers serve as the foundation to improve the validity of collected confidence data. Recommendations are presented to improve the clarity and utility of confidence tiers to provide meaningful evaluation of target constructs.

Keywords: Confidence tier, Metacognition, Self-evaluation, Assessment design

Introduction

For over a century, assessment developers have used confidence tiers to prompt assessment respondents to evaluate their own knowledge. This practice has been used in psychology, social sciences, and increasingly in discipline-based education research (DBER) fields to weight assessment scores, evaluate degrees of metacognitive knowledge, and make assumptions regarding the strength of alternate conceptions in research-based assessments. Since its first mention in the literature in the early 20th century, this practice has been referred to by various names (e.g. confidence testing or self-reflective prompting), though in assessment contexts, the use of this practice as a second tier for a knowledge/content item has been referred to as a confidence tier. The target construct of the confidence tier has been vaguely defined due to its long history and expansive applications across disciplines, resulting in challenging ambiguity for assessment developers to meaningfully operationalize this tier.

In an assessment climate which prioritizes validity and reliability evidence for its measures, the construct for this practice must be clearly specified to provide meaningful measurement. This work leverages response process validity interviews with undergraduate students to explore participants' thought processes with confidence tiers in varied formats. The purpose of this study is to investigate how learners interpret and engage with different features of confidence tiers across assessment items. *Through rich qualitative analysis, we seek to provide specific recommendations for assessment developers to collect more accurate data and draw more specific and operational conclusions.*

Background

Uses of Confidence Testing – Weighted Models

Literature detailing the use of confidence tiers spans to the early 20th century where it was described as a method for increasing the amount of information that could be elicited through multiple choice assessments (Echternacht, 1972; Henmon, 1911; Hollingworth, 1913; Trow, 1923). Early proponents of this technique were clear in its limitations, describing that "*while there is a positive correlation on the whole between degree of confidence and accuracy, the degree of confidence is not a reliable index of accuracy*" (Henmon, 1911).

In early applications, confidence testing reportedly improved the reliability of an assessment by weighting responses based on the examinee's degree of confidence (Hevner, 1932; Soderquist, 1936). Weighting examinees' responses were subsequently used to correlate personality factors with assessment behaviors (Swineford, 1938; Wiley & Trimble, 1936; Ziller, 1957), and to explore how partial knowledge could be assessed (Archer, 1962; Coombs, 1953; Coombs et al., 1956).

A notable shift was made towards using statistical decision theory and the concept of probability in confidence investigations, accounting for partial knowledge as well as the introduction of an "easing-in" process where respondents were slowly introduced to the idea of attaching a quantity to their personal probability beliefs (de Finetti, 1965). This approach allowed psychometricians to consider the contribution of psychological factors in probabilistic testing, such as the relationship between the degree to which examinees display confidence in their responses and certain personality traits (Ebel, 1965; Hansen, 1971).

From early weighted-model investigations, examinees were found to possess a characteristic tendency to be either certain or uncertain which cannot be fully accounted for by stability of knowledge (Hansen, 1971). These individual differences in personalities threaten the validity of the data produced by weighted-score techniques (Jacobs, 1971) and have broadly led to declined use of weight-score techniques in favor of investigating the role that metacognition plays in the retention of learned material.

Uses of Confidence Testing – Metacognition

Metacognition, often defined as "knowing about knowing" (Flavell, 1979; Garner & Alexander, 1989), is broken down into two components, where the *cognitive knowledge* a learner has regarding their own thinking processes directly impacts the degree of *cognitive regulation* that must be done to monitor and evaluate their processes of thinking (Jacobs & Paris, 1987). The construct of *metacognitive calibration* is defined as the level of agreement between a respondents' metacognitive knowledge and their performance on a cognitive assessment (Keren, 1991; Yates, 1990) which can be used to draw conclusions about the relationship between learners' metacognitive regulation and cognition.

Dunning & Kruger (1999) provided evidence of a phenomena termed *metaignorance*, where the most "unskilled" in a domain are most likely to be "unaware" of their degree of competency, compared to those with higher skill. This representation of miscalibration in low-performing examinees has gone on to be cited in thousands of publications across disciplines, frequently in educational contexts. Assessing for a lack of

metacognitive regulation skill (or metaignorance) has increasingly become a prevailing motivation for modern work in assessment across diverse fields. Within classroom settings, education researchers have cited the prevalence of metaignorance to cluster students and target those with low metacognitive calibration for further tailored instruction (Bell & Volckmann, 2011; Nietfeld et al., 2005; Pazicni & Bauer, 2014), or provide specific training to improve metacognitive regulation and learning outcomes (Lavi et al., 2019; Thiede et al., 2011; Zimmerman et al., 2011). In professional certification programs, the degree of a respondent's metacognitive calibration has been built into dynamic algorithms to tailor which content questions they must revisit in the future (Sun et al., 2016).

Developing metacognitive regulation skill has been shown to improve learners' assessment performance and study regulation (Thiede et al., 2003) and has been linked to high performance on problem-solving tasks (Swanson, 1990). In STEM-specific studies, improving metacognition has been found to improve outcomes in science literacy, learning, and teaching (Adey et al., 1995; Blank, 2000; Davis, 1996; Georghiades, 2000, 2004). Increasingly, the literature indicates that explicit metacognitive regulation can improve performance in STEM courses (Bunce et al., 2023; Casselman & Atwood, 2017; Cook et al., 2013; Dori et al., 2018) and that implicit metacognitive monitoring does not improve learning performance alone (Hawker et al., 2016). These findings have reinforced the need for explicit metacognitive prompting, contributing to the profusion of confidence tiers in research-based assessments over the past fifty years.

Multi-Tiered Partial Knowledge Assessment Models

As cognitive theorists worked to propose different explanations for how knowledge develops in the minds of learners, educational measurement scholars have sought to develop assessment tools which measure formal reasoning ability across large groups of learners (Burney, 1974; Lawson, 1978; Staver & Gabel, 1979). This assessment approach evolved into two characteristic tiers (answer and reasoning) intended to measure partial knowledge.

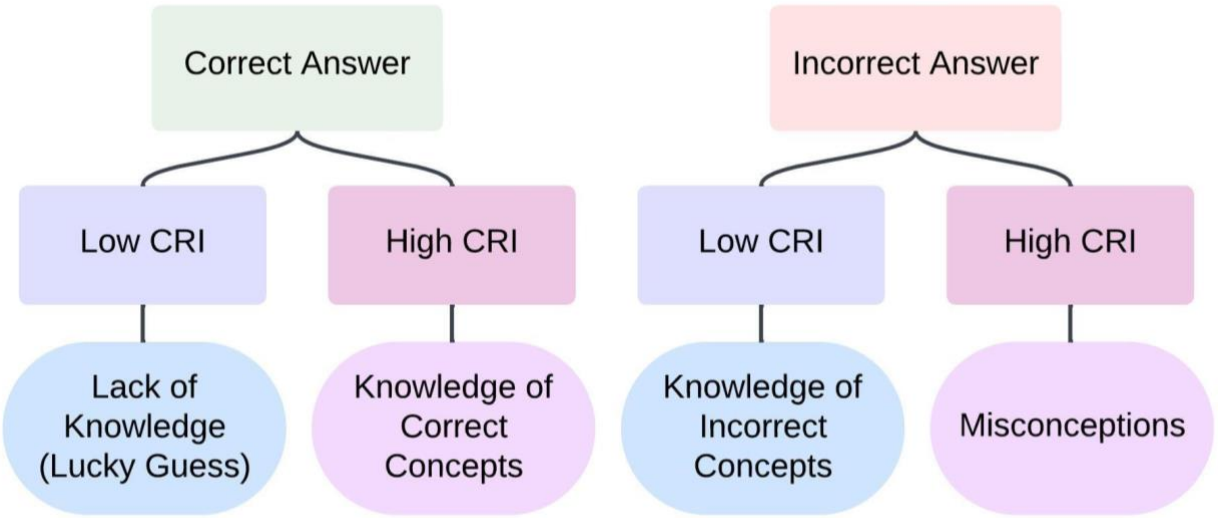
Treagust's seminal work (1988) used this structure as a "diagnostic test" to evaluate alternate conceptions (ACs) by asking STEM students to indicate a correct answer and then select a reasoning for why their response was correct in a secondary tier (Treagust, 1988). This ushered in a well-known class of assessments called concept inventories (CIs), designed to elicit whether students have scientifically appropriate ideas within a particular context, i.e. Newtonian beliefs about force (Hestenes et al., 1992).

In CI development, confidence tiers have been used as a metric of partial knowledge in place of a reasoning tier. A key demonstration of this technique is Hasan et al.'s (1999) assessment of classical mechanics where the strength of ACs was determined by the degree of certainty in their selected answer. This application uses the Certainty of Response Index (CRI) to facilitate an assessor's identification of items a student answered incorrectly due to a strongly held AC or were answered correctly due to guessing. This identification practice (the "Hasan hypothesis") provided a decision matrix for educators to identify the presence of ACs (Figure 1) based on self-reflection of their content knowledge (Hasan et al., 1999). The Hasan hypothesis represents a key pivot in DBER literature towards the use of a confidence tier to gauge the presence of ACs.

Three- and four-tier assessment formats of an answer and reasoning tier combined with one or two confidence tiers have also been used as hybrid diagnostic tests (Al-Rubayea, 1996; Caleon & Subramaniam, 2010a; Caleon & Subramaniam, 2010b; Franklin, 1992; Hill, 1997). Despite increased sensitivity, the limitations of four-tiered assessment structures support use of a restricted two-tiered answer and confidence format (Abell & Bretz, 2019; Brandriet & Bretz, 2014; Connor et al., 2021; McClary & Bretz, 2012) or keeping a three-tier format of an answer, reasoning, and confidence tier (Assimi et al., 2024; Liampa et al., 2019; Taslidere, 2016; Yang & Sianturi, 2019). As assessments using confidence tiers have become more

commonplace, the stems and scales used to measure the construct of confidence have varied, prompting a further look at the specific design features and interpretability of this format.

Figure 1. Decision tree based on combinations of correct or incorrect and low or high CRI, adapted from decision matrix in Hasan et al., 1999



Confidence Tier Design Features and Target Usage

As the use of confidence tiers has evolved and changed over time, so has the specific language used in different item parts. The prompt of the confidence tier (*stem*) provides the direction and context for the respondent. Examples of different stems and their source publications are included in Table 1. While not exhaustive, this list of examples demonstrates the variation in prompts used as confidence tier stems. Some use specific terminology referring to "confidence", while others focus on certainty, correctness, or the type of information used to answer an item.

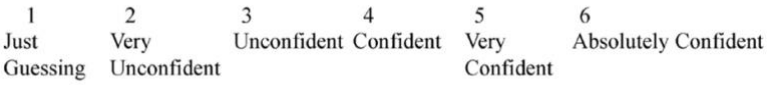
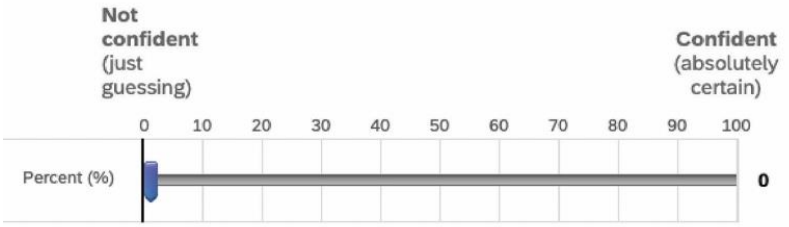
Table 1. Example stems from studies using confidence tiers

Stem	Publication
How sure are you?	Hunt, 2003
For each answer you select, indicate whether you answered using knowledge learned in classes/books or not.	Hasan et al., 1999
Indicate your confidence that your answer to the previous question was correct.	Lundeberg et al., 1992
How well do you think you understand the text you just read?	Griffin et al., 2008
How confident are you about the answer you chose?	McClary & Bretz, 2012
For me to X, it would be… (on a scale of very easy to very difficult)	Ajzen & Madden, 1986

The other component of the confidence tier is the range of options (*scale*) provided for a respondent to choose from. Table 2 provides several examples of the scales used in confidence tier publications which illustrate the variation in the terminology and the inclusion of numeric and descriptive response options. Another notable difference is in the number of options included (even or odd-numbered) and the presence or omission of "average." Many scales have each response labeled, however, others (i.e. Trafimow et al., 2002; Webb et al., 1994) are made up of anchor point descriptions with unlabeled options to represent

equidistant degrees of agreement. One example (McClary & Bretz, 2012) features two endpoint anchors with several unlabeled response options between them.

Table 2. Example scales from studies using confidence tiers.

Scale	Publication
○ Sure ○ In Doubt ○ Guessing	Swineford, 1938
1. The answer is probably true 2. The answer is possibly true 3. I have no basis for response 4. The answer is possibly false 5. The answer is probably false	Ebel, 1965
1. Not Certain 2. 3. Somewhat Certain 4. 5. Absolutely Certain	Webb et al., 1994
0. Totally guessed answer 1. Almost a guess 2. Not sure 3. Sure 4. Almost Certain 5. Certain	Hasan et al., 1999
	Händel & Fritzsche, 2015
	Caleon & Subramaniam, 2010a
1. Not at all easy 2. 3. 4. Extremely easy	Trafimow et al., 2002
	McClary & Bretz, 2012

Many confidence tiers feature Likert scales to gauge degree of agreement or alignment to the construct of "confident vs not confident". Prior work on Likert-style scales has suggested that responses corresponding to average or midpoint option may indicate a null response or "dumping ground" (Chyung et al., 2017; Kulas et al., 2008), be related to social desirability, or linked to certain personality traits (Garland, 1991; Hernandez et al., 2018). In practice, respondents prefer scales which include a midpoint (McDonald,

2013; Preston & Colman, 2000), however, Simms et al. (2019) found no significant difference between odd- and even-numbered scales psychometrically.

When measuring latent constructs with Likert scales, dichotomizing or trichotomizing responses during analysis can be performed without sacrificing validity and reliability metrics (Matell & Jacoby, 1971). Considering that extreme options are more frequently selected (Albaum, 1997), increasing the number of response options on a Likert scale may offer negligible advantages. For the optimal number of options in a scale, precision decreases clearly below 4 and does not strongly increase after 7 options (Lee & Paek, 2014; Lozano et al., 2008).

Another frequently used scale is the visual analog scale (Aitken, 1969; Flynn et al., 2004; Hayes & Patterson, 1921; McClary & Bretz, 2012), which appears as a horizontal line with two opposing anchors on either end (e.g., agree vs disagree, confident vs not confident). Respondents indicate their position on the interval with a slider bar to provide a measure of the construct. There is modest evidence of psychometric superiority for this method over Likert scales (Hilbert et al., 2015; Russell & Bobko, 1992), though they have also been demonstrated as statistically similar (Bergman, 2009). Online data collection has made the use of analog scales more prevalent, although increasing the number of response options beyond 7 offers diminishing returns (Simms et al., 2019).

To develop measures which investigate self-regulated learning, "easy-difficult" scales have been used citing evidence that these scales investigate the same construct as a "confident-unconfident" scale (Manstead & van Eekelen, 1998; Sparks et al., 1997; Terry & O'Leary, 1995; White et al., 1994). This differentiation in scale design to evaluate academic self-reflection has been evaluated quantitatively, but investigation into learners' interpretation of this scale would offer clarity for assessors who are looking to qualitative literature to make design decisions.

As evidenced in this survey of the variations in assessment design, the target construct of a confidence tier has become increasingly difficult to characterize as its uses have expanded. In many cases, the confidence tier has been leveraged to examine the strength of alternate conceptions in learners, although it is also widely used to investigate and measure metacognition. Despite the substantial differences in these research goals, the same tool (confidence tier) is used to collect data and draw conclusions. Determining the nuance in how respondents interpret features of confidence tier stems and scales is therefore critical to design more accurate measures.

Recent work has identified the broad range of factors that learners consider when ranking their confidence on research-based assessments in a DBER context (Koevoets-Beach et al., 2023), laying groundwork for qualitative exploration into how respondents interpret confidence tiers to help tailor stems and scales for the desired outcome of the assessment. This current study seeks to build on over a century of confidence tier literature to characterize respondents' interpretations of different stems and scales and help provide specific recommendations for future use of this technique in education research.

Framework

This research is framed by social cognitive theory (SCT; Bandura, 1977, 1986) and the theory of self-regulated learning (SRL; Zimmerman, 2000, 2002).

SCT outlines learning as a practice which is affected not only by an individual's internal personal factors (i.e. cognition and emotions), but by the reciprocal interactions between personal factors, an individual's behaviors, and their external environment (Bandura, 1977, 1986). These three categories of factors (personal, behavioral, and environmental) are mutually interactive and inform the overall experience of learning. Under SCT, it is assumed that when learners engage in assessments, they are not just cognitively engaging with the

assessment items through their content knowledge, but also through their environmental and behavioral experiences around assessment and the content area. Using this holistic view of cognition allows for qualitative analysis of respondents' engagement with confidence tiers to expand beyond the item itself and looked at as a piece of a larger cognitive picture. Within SCT, Bandura (1997) identified general components of SRL to include the concepts of metacognition and motivation alongside cognitive processes. SRL (Zimmerman, 1989, 2000, 2002) narrows in on the cyclical process of learning.

SRL theory outlines three phases in the learning cycle: forethought, performance, and self-reflection. Forethought occurs prior to learning and involves goal setting and planning, as well as motivational beliefs such as self-efficacy and interest. The performance phase includes strategies learners implement during learning, such as note-taking or consulting a textbook, as well as monitoring the progress of learning while implementing these strategies. The last phase is self-reflection, which includes evaluating one's performance, adapting strategies to enhance performance, feelings of success or failure, and a sense of satisfaction regarding the learning task. Learners vary in the degree to which they self-regulate, and some may demonstrate more well-developed self-regulatory skills than others (Ning & Downing, 2015; Yip, 2007). The SRL cycle describes the thoughts, feelings, and behaviors that learners exhibit while accounting for personal factors associated with academic resilience, such as goal-setting, self-efficacy, control over study efforts, and motivational beliefs (Borman & Overman, 2004; Freeman et al., 2004; Miller, 2002). This work uses SCT and SRL complementarily to explore how students respond to, engage with, and interpret confidence tier stems and scales as tools for self-reflection and metacognition within an assessment setting.

Research Objectives

This study aims to explore how learners interpret and respond to a range of stems and scales used in confidence testing. This work also explores whether the way these components are presented influence how learners think about and respond to questions. Therefore, the two research objectives for this study are:

1. Investigate learners' interpretations of different stems and scales through their engagement with confidence tiers.
2. Generate recommendations for future administrations of confidence tiers.

Methods

The study design and participant recruitment were approved to meet ethical human subjects research standards by an Institutional Review Board (IRB#: 22.0454).

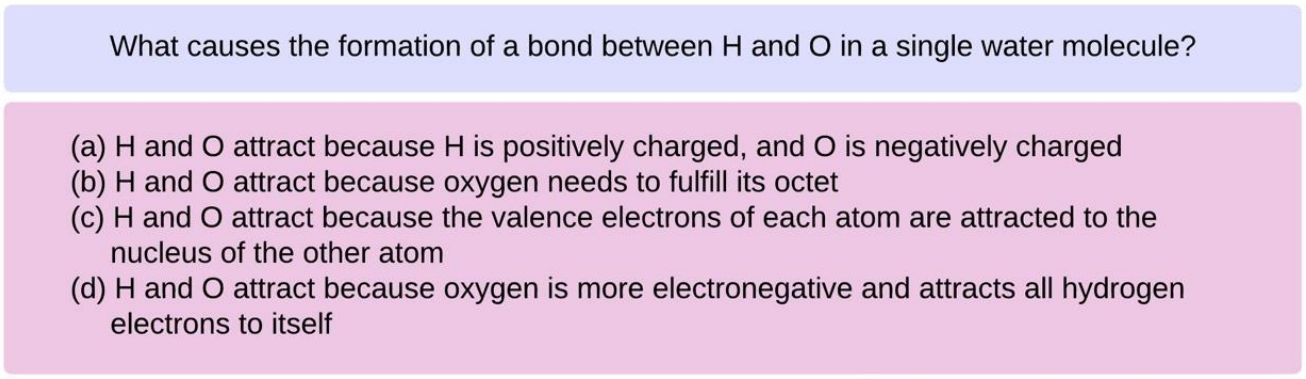
Participants

Participants for this study were ten students at a mid-sized, public university in the southeastern United States. This study was conducted using items from a research-based assessment of General Chemistry concepts, so participants who had completed the two-semester General Chemistry course sequence within the last year were eligible to participate. Participation in the study was voluntary with participants recruited from another related research project and compensated for their time with a gift card. All participants were provided the informed consent form to review and sign prior to the interview and consent was verbally described in person prior to the initiation of the interview. Participants were assured of the confidentiality of their responses, and measures were taken to anonymize the data by assigning pseudonyms.

Data Collection

Semi-structured interviews were conducted during the Fall 2023 and Spring 2024 academic semesters, with interviews continuing until saturation had been reached. The interview protocol was designed to explore participants' ideas, perceptions, and overall knowledge related to the target course content. The interview was contextualized in General Chemistry to provide some grounding for students to evaluate the range of stems and scales. Interviews consisted of three phases. In the first phase, participants were first presented with 2-3 content questions (example in Fig 2) from a research-based assessment of General Chemistry knowledge (Balabanoff et al., 2022) and asked to express their thought processes aloud while responding. They were prompted to identify the target concept for each question and reflect on their feeling of correctness for their selected answer. This initial exercise was used as an "easing in" process (de Finetti, 1965) to help participants feel comfortable and engage in recall of target General Chemistry concepts. During this phase and over the course of the interviews, participants were asked reflection questions to gauge their confidence and capture their understanding of broader course material.

Figure 2. Sample content question from research-based assessment used for interview content.



After engaging with single-tiered assessment items, the second phase of the interviews directed students to engage with five distinct confidence stems, each addressing different aspects of confidence and understanding. These stems (Figure 3) were adapted from existing literature to align with different techniques that have been used to capture students' confidence and metacognition. The phrasing was intentionally varied to allow for comparisons to be made between different stem verbiage. Participants were prompted to discuss their interpretations of each stem and discuss what type of reflection they would engage in based on the language used; then indicated which stem they felt best prompted self-reflection on their content knowledge. This process was repeated with five different response scales which consisted of descriptive Likert-style rating scales with both odd and even numbers of response options, a visual analog scale, an adaptation of Hasan's CRI, and an 'easy-difficult' scale (Figure 4). Students were asked to engage with each scale option, discuss their interpretations and comparisons of each, then indicate which they felt best captured confidence in a response. The order of these stems and scales was the same for each participant but varied to avoid the appearance of fixed pairs.

The third interview phase prompted students to use their individually selected stem and scale with a new content item to simulate a novel two-tier format. This approach allowed participants to personalize their self-assessments and allowed them to reflect on their engagement with the confidence tier, providing insight into the functionality and interpretation of the stem and scale they felt best prompted and captured their self-reflection. The range and frequency of selected options are illustrated in Figure 5 and these selections provided structure for the qualitative coding process as respondents' reasoning and interpretations of the stems and scales were inductively coded.

Metrics such as cognitive load, impact of item content, and extended response time associated with confidence tiers have been investigated in past literature (see Background) but were not measured within the

Figure 3. Confidence tier stem options used in interviews

Stem 1	How well were you able to answer the question above?
Stem 2	How well do you feel you understand the concepts being asked about?
Stem 3	How confident are you in your responses to the question above?
Stem 4	How much did you use knowledge learned in class to answer the question above?
Stem 5	How would you describe the difficulty of the question above?

Figure 4. Confidence tier scale options used in interviews

Scale 1

- Very Well
- Well
- Poorly
- Very Poorly

Scale 3

0	Totally Guessed Answer
1	Almost a Guess
2	Not Sure
3	Sure
4	Almost Certain
5	Certain

Scale 2

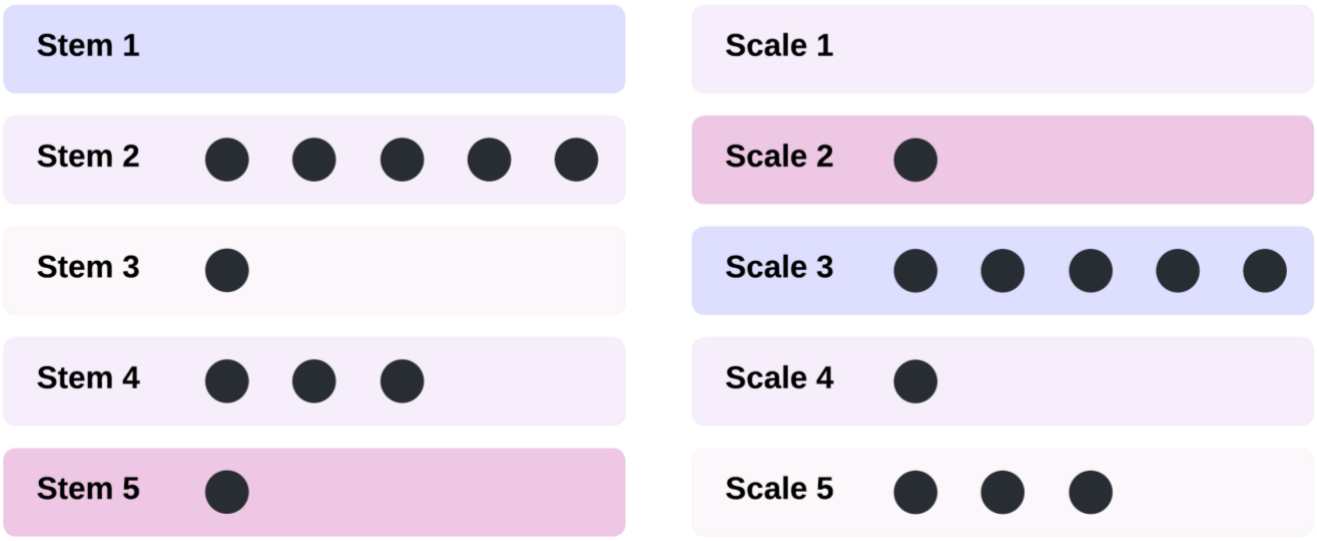
- Very Well
- Well
- Average
- Poorly
- Very Poorly

Scale 4

- Extremely Difficult
- Moderately Difficult
- Slightly Difficult
- Slightly Easy
- Moderately Easy
- Extremely Easy

Scale 5

Figure 5. Frequency distribution of each selected stem and scale in the third interview phase



scope of this work. This is because previous metacognitive calibration studies weighing these metrics found that the two-tier format optimizes metacognitive benefits while balancing the burden of test-taking time and cognitive load (Caleon & Subramaniam, 2010a; Caleon & Subramaniam, 2010b). This supports the current study's focus on the design, phrasing, and functionality of items, using interview data to provide an in-depth description of students' connections between confidence and comprehension.

Data Analysis

Interview data was cleaned through an online transcription service (Otter.ai) and analyzed using qualitative coding software, Dedoose® Version 9.2.6, which provided a structured approach for recognizing and categorizing themes based on student responses. Each interview was inductively coded beyond which specific stems and scales were selected by participants (Miles et al., 2014). This allowed for the sequential identification of thematic codes that emerged regarding students' interpretations of confidence, understanding, and the assessment process.

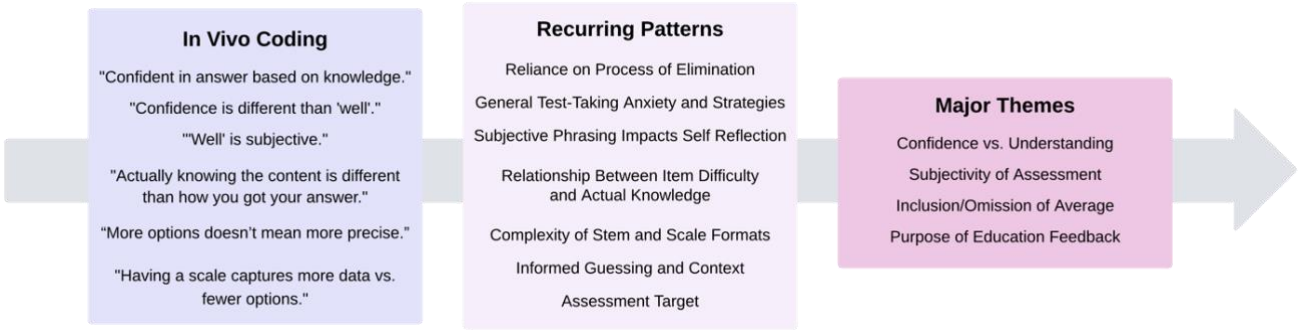
The coding process began with in vivo coding, which used words and short phrases from the participants' own language in the interviews as codes (Miles et al., 2014). This allowed for pattern recognition of repeated phrases by the participants. The initial set of codes were aligned with the two research objectives and the range of stems/scales provided to students, which provided a broad spectrum of themes that reflected the diversity of student experiences and interpretations.

As coding progressed, patterns began to emerge across the interviews, providing insights into how learners perceive and express confidence and understanding within the context of a research-based assessment. Through the iterative coding process, emergent pattern codes were developed (Miles et al., 2014). Initially, the data were coded into specific categories, but as the analysis deepened, overlapping, redundant, or overly broad codes were merged to create more cohesive themes. For example, the distinct codes "Confident is subjective" and "Well is subjective" were eventually combined into a single code reflecting the shared underlying concept of subjectivity in self-assessment. This approach followed the iterative methods aligned with constant comparative analysis (Glaser, 1965), which emphasizes the importance of ongoing code refinement to capture the complexity of qualitative data.

To ensure consistency and reliability in this process, two coders independently reviewed the interview data and applied initial codes. After the independent coding phase, coders met to compare their

interpretations, revisit and revise any discrepancies, and reach consensus on the final set of codes. This structured approach ultimately led to the identification of four major themes that encapsulated the core aspects of the students' responses (Figure 6).

Figure 6. Process of qualitative coding process and identification of major themes



Findings

Participants' engagement with confidence tiers fell into four major themes: "Confidence" *vs* "Understanding", Subjectivity of Assessment, Inclusion/Omission of Average, and Purpose of Educational Feedback (RO1). These themes elucidate the features of confidence tier stems and scales which allow for precise prompting and measurement within a two-tier assessment (RO2).

RO1: Investigate learners' interpretations of different stems and scales through their engagement with confidence tiers

Theme 1: "Confidence" vs "Understanding". When engaging with various items, students distinguished between their "confidence" in an answer and their "understanding" of underlying concepts. From participants' descriptions, "confidence" results from test-taking skills or familiarity with phrasing. While deep comprehension of a subject may increase confidence, it is complementary to an underlying feeling of test-taking competence. This sentiment echoes previous confidence tier literature which has indicated that confidence tier responses are often based on underlying personal factors (Ebel, 1965; Hansen, 1971; Koevoets-Beach et al., 2023).

Participants described that conceptual knowledge needs to be comfortably applied within an assessment for a respondent to rank high confidence in their response. For example: *"If I understand the concept, but I don't feel confident in the answer, that has to do with something on the application step",* indicating a cognitive division between the two constructs where one could exist without the other (*e.g.* deep conceptual understanding with low confidence due to unfamiliar phrasing/terminology). Participants interpreted "understanding" prompts (i.e. Stems 2 and 4) as probing their metacognition of content knowledge and interpreted "confidence" prompts (i.e. Stems 1 and 3) as probing their test-taking skill.

During the third interview phase, participants primarily selected "understanding" options for their individual two-tier design, citing more critical self-assessment. Participants described evaluating cognition more honestly with these "understanding" stems: *"Even if I can get the answer, if I'm not able to understand how I'm getting [it], I'm doing worse off because I'm not actually understanding those concepts".* Participants described deeper metacognitive reflection between learning and applying when responding to "understanding" stems: *"If I were to say that I applied knowledge that I learned in class but my confidence level was low, maybe the learning in class is not translating to being confident about using it".* This distinctly contrasted engagement with "confidence" stems, where participants frequently defaulted to time dependency and/or test-taking strategies: *"I always feel confident my answer at the end no matter what because I use process of elimination".* "Confidence" stem responses often

circumvented metacognition by determining how quickly they could respond or how many multiple-choice options they could eliminate.

This indicates that using more specific language regarding "conceptual knowledge" in place of "confidence" in stems may be more effective in engaging respondents in and promoting the practice of metacognition. Stems which direct attention to the learning process and application of understanding helps learners focus on deeper metacognition rather than surface-level self-evaluation.

Theme 2: Subjectivity of Assessment. Respondents possessed a strong understanding of the subjective nature of assessing one's knowledge, as each individual's definition of terms such as "well" or "confident" may vary enough to muddy responses: *"It really depends on what the word 'well' means…if 'well' means 'correctness' or 'do you understand what the question is asking?' or like, 'can you recall that this is something that you've learned before?' And if I can do all three, I answered it very well. But if someone else were to see this, maybe that definition could change."* The subjectivity of phrasing was frequently cited by respondents when they described engagement with different stems.

When engaging with subjective terminology, respondents struggled to appropriately reflect on how they felt about their performance: *"[Stem 1] is a very open question because you don't get a lot of context for my answer. I would like it better if it was more specific"*. In contrast to that, participants using stems and scales that specifically captured guessing and/or uncertainty felt that rankings closely matched their own feelings. For example, in reference to Stem 2, one participant described: *"If I'm having to describe how well I* understand *the concept, I'm gonna get a better idea of how I actually did"*. This supports the use of stems that directly prompt metacognitive reflection rather than test-taking ability to redirect focus to the self-regulation step of the SRL cycle.

Participants emphasized the complexity of assessing understanding: *"It can be just overwhelming... it's hard to quantify confidence,"* but felt that the subjectivity of ranking could be mitigated through the inclusion of a complementary number scale, like in Scales 3 and 5. This was described as providing both scope and specificity to one's rankings: *"I feel like I would fall in the middle but with [Scale 3] there are more options, you get a numerical answer. I think combining words and numbers makes the most sense to think about if you actually knew it or not."* Participants found both the greater resolution of possible options and the addition of a numeric range (as in Scales 3 and 5) to be helpful in conceptualizing their confidence: *"Numbers are easier to quantify but having something like this where the scale has both numbers and words gives more room for somebody to be like 'I'm not like doing bad, but I'm not doing great'"*. The ability to select from a combination scale offered both rich description and specificity for students to rank their degree of certainty in their answer.

The utility of a full 1-100 scale (Scale 5) was questioned by some respondents, who described it as overwhelming and increasing the cognitive load required to engage in each individual confidence tier: "*when there's a lot of options, that doesn't really make it more precise, it just adds more options"*. While several respondents felt the large scale helped mitigate the effects of subjectivity, others characterized Scale 5 as having the largest degree of subjectivity due to the differences in individual scoring practices: "*every student has a different level of "confident" in their brain. For one person it could be a 70, and for another it could be a 90*". To strike a balance in descriptive power and specificity of construct, a constrained numerical scale may be most useful for prompting metacognition.

Theme 3: Inclusion/Omission of Average. During interviews, participants were asked to describe the effects of the inclusion of the "average" midpoint on a confidence scale. Participants felt that an average option captured their uncertainty more effectively. Use of a midpoint option was frequently cited as helping to prevent under- or overestimating one's abilities and allowing for more balanced self-evaluation. When presented with an even-numbered scale without a neutral option, students stated they would likely rate themselves on the lower end of the scale: *"I would probably put 'poorly'. Just because I would not say that I was confident enough to say 'well'."* Participants described this as a shortcoming for even-numbered scales, though

engaging in difficult self-reflection may still be productive for redirecting attention and study efforts if the target is metacognitive regulation.

When participants employed odd-numbered scales with a midpoint, we observed its use as a "dumping ground" for unsure responses (Kulas et al., 2008): *"If I'm not sure about it […] I just prefer using 'average' rather than having to put myself on one side or the other"*. This avoidance of ranking by relying on the neutral response could inhibit meaningful reflection and measurement. Respondents also chose average when they had used some content knowledge in conjunction with test-taking strategies but were still unable to confidently select a correct response "*I started with the process of elimination but […] since there doesn't seem to be a middle ground, I wouldn't know where to put myself."* The selection of average reflected learners' ability to partially navigate a question but lacked the specific language an assessor would need to identify why a respondent ranked their confidence as average.

Participants indicated that while "average" might help in comparing oneself to others, it doesn't provide clear, direct information about personal understanding: *"If they [respondents] are going to compare themselves...you'll get more of that [comparison] and less of the topic itself."* Some provided their perspective for an assessor concerning assessment data output quality: *"I think when you're looking at data and trying to compare results... if you have a bunch of average answers, then you don't really know as much."* Offering an even-numbered scale may force individuals to assess their performance more critically, as one participant noted: *"not having an average would probably be better because you're forcing someone to decide whether they think they did well or not."* While the inclusion of average is a well-documented preference for survey respondents (McDonald, 2013; Preston & Colman, 2000), it is important for assessors to consider whether prompting more critical self-reflection is important in their assessment context. Constraining respondents to select either high or low confidence or using specific language that captures "informed guessing" or "feeling unsure" may promote more targeted metacognitive engagement and provide richer data for analysts.

Theme 4: Purpose of Education Feedback. When presented with different stem and scale options, participants reflected that the intention for using the confidence tier was clearly tied to both learners' and assessors' goals: *"I think whoever's reading this [stem] would be able to gauge their learning better. If they are having to describe how well they understand the concept, then you [the assessor] are gonna see how good or bad they actually do"*. Participants valued the confidence tier for prompting reflection on their understanding of the underlying concepts while acknowledging that the tier could be helpful in a classroom setting to identify ACs: *"I think [Stem 2] would help students in the long run because their answer then makes the professor realize that they don't understand a certain concept"*. Respondents felt the data generated could allow instructional staff to revisit alternate conceptions, benefiting students' overall comprehension. Specifically, they felt that prompting students to gauge their understanding of concepts provided the most meaningful information to the assessor: *"You want to know if somebody's guessing or knowing the content. I feel like that's the point of including the reflection question."* This highlights students' awareness that engagement with these stems prompts their own metacognitive reflection while also providing valuable insights for their instructors. They also identified features that would be unhelpful for an assessor, such as Scale 4 prompting for a difficulty ranking: "*whether or not something was difficult or easy for someone would be less important for the teacher just because it could just be reflective of how well you studied or sometimes just like if you got enough sleep*". Overall, students felt that prompting reflection on difficulty would be unhelpful both for the respondent and the assessor.

Participants contrasted the criteria for features most helpful for respondents to what would best serve assessors. For example, having more scale options (as in Scale 5) allows more breadth for learners' accurate assessment, but constraining stem options could help assessors interpret the assessment data better: "*If you have fewer [scale] options, you kind of get people pushed to extremes so you're able to get a broader range of data*." This idea of differing priorities was also seen in the discussion of including an average or midpoint on a scale (see Theme 3). The inclusion of average for their own self-reflection helped them capture neutral feelings despite

potentially limiting an assessor's interpretation of students' data. Acknowledging how the same feature (i.e. a neutral option) functions for different stakeholders, respondents exposed a conflict between an assessor's collection of meaningful measurements with a learner's ability to engage in self-reflection and metacognitive regulation. This conflict is mirrored in the varied historical use of confidence tiers and the ability of an assessor to make claims about the validity of the data collected using these tools.

RO2: Generate recommendations for future administrations of confidence tiers

To develop recommendations for assessment developers who seek to use confidence tiers, the four themes observed in students' response processes have been used to help tailor stems and scales to best target constructs of interest. Considering these findings and existing confidence tier literature, there is a distinct separation between two target constructs for the use of confidence tiers: (a) engaging learners in metacognition, and (b) measuring self-evaluation of conceptual knowledge (Table 3). While complementary, each should be explicitly targeted to provide the most clarity for respondents and meaningful data for assessors.

Table 3. Definitions and examples of two target constructs of confidence tiers

Target Construct	Definition	Focus	Example from Interview Data
Metacognition	The *practice* of reflecting on the degree (or strength) of one's knowledge of the content being assessed	Respondent	*"I think whoever's reading this would be able to gauge their learning better."*
Self-evaluation of Conceptual Knowledge	The *measurable ranking* of one's confidence or certainty in the correctness of their answer	Assessor	"*You [the assessor] want to know if somebody's guessing or knowing the content. I feel like that's the point of including the reflection question."*

Engaging students in metacognition. For confidence tiers targeting students' evaluation of their content knowledge, our findings indicate that using language which directs attention to the *understanding* of concepts more effectively activates metacognitive regulation. Stem phrasing targeting a student's "confidence" or "ability" opens their interaction with the confidence tier to consider test-taking strategies, personal self-efficacy beliefs, and time dependence. Therefore, for a metacognitive confidence stem, learners should be prompted to reflect on the "strength" or "degree" of their understanding.

To design a metacognition-targeted confidence scale, a key consideration is respondents' use of a mid-point option. Unlabeled midpoint options or "average" labels can lack specificity and are more likely to become a "dumping ground" for unsure responses or social desirability bias (Garland, 1991; Kulas et al., 2008). Our data supports the use of odd-numbered response scales which shift away from "average" towards "informed guessing" or "feeling unsure" to allow learners to extract more meaning from their own results and better direct their study efforts. If using a complementary numerical scale, a constrained number of options (e.g., Scale 3) strikes balance between descriptive power and specificity.

For metacognitive regulation of study efforts, gradual and consistent exposure in the form of metacognitive training has been recommended (Papaleontiou-Louca, 2003; Pazicni & Bauer, 2014; Zimmerman et al., 2011). Confidence tiers used to improve metacognitive regulation should therefore be implemented with formative assessments which are themselves designed to provide frequent feedback to

learners about their learning processes (Harlen & James, 1997). By pairing confidence tiers with formative assessments such as clicker questions or checkpoint quizzes (Bunce et al., 2023; Colthorpe et al., 2018), assessors provide opportunity for learners to engage in self-reflection and comparative benchmarking with peers. For assessments designed to be summative, confidence tiers may be less impactful as a metacognitive tool for learners but still offer valuable information to assessors to measure the strength of alternate conceptions.

Measuring self-evaluation of conceptual knowledge. Recommendations for designing confidence tiers to specifically target the strength of alternate conceptions requires further consideration of how their output impacts assessors' interpretations. As confidence tiers are operationalized differently, design features follow. Specifically, confidence tier stems targeting measurement may benefit from more narrowly targeting the construct of interest. For example, an amendment to the stem used for our General Chemistry assessment may prompt students: "*How certain do you feel you understand the chemistry concepts asked about above?*" Measurement of learners' understanding of a certain concept may be conducted in the context of either a summative or formative research-based assessments.

For confidence tier scales, a major design feature to consider is the inclusion of an average/mid-point option or the use of an even number of options within a scale. Midpoint or "average" options allowed respondents to convey uncertainty about their performance and to avoid providing a negative assessment when unsure. For assessment in high-attrition courses such as General Chemistry (Koch & Drake, 2018) where a static negative self-reflection may be pervasive (Atherton, 2017), providing an average option may help mitigate bias in self-evaluation and allow for a more accurate measurement. Respondents report a preference to have a midpoint for their own self-reflection, but acknowledge the likelihood for it to become a sink for unsure responses, potentially inhibiting the transmission of helpful information to an assessor. We recommend using a scale with no midpoint or using the specific terminology of "made an informed guess" rather than "average", "neutral", or no label to help direct attention to self-regulatory knowledge. Participants recognized that restricting the number of scale options and eliminating average could provide more meaningful data output for an assessor to use to improve instruction and learning outcomes, aligning with existing literature (Lee & Paek, 2014; Lozano et al., 2008; Simms et al., 2019).

Participants in our interviews described that confidence scales which combine numerical values and descriptions could help respondents interpret the thresholds more consistently with one another. By combining these two features (similarly to Scale 3) assessors can provide more clarity for the student while controlling for the subjectivity identified in a large numeric scale with only two anchor labels. This also helps account for the problem of ordinality associated with Likert-style scales (Jamieson, 2004), specifically for those who seek to quantitatively analyze assessment responses as interval data.

Using a scale which clearly identifies uncertainty and informed guessing helps assessors identify alternate conceptions and pinpoint where extra instruction and assistance could be directed. Scale phrasing should allow respondents to identify these differences. We recommend using a restricted numeric scale (4-6 options) alongside highly descriptive Likert-style responses to allow for assessors to make stronger claims regarding the unidimensional and interval qualities of their measurements and draw clearer conclusions about conceptual understanding.

Implications

Assessment developers have used confidence tiers over the last century to investigate students' cognitive models and metacognition. For assessment to be meaningful to both assessors and students, the purpose of the assessment must be clearly defined and aligned with an understanding of the construct being measured. Much of the murkiness that has been demonstrated in the output from confidence tiers has stemmed from

the misalignment with their intended purpose. Extant literature supports using confidence tiers to assess and improve learners' metacognitive regulation to improve learning outcomes, therefore assessors must (1) understand the construct being measured, (2) select the outcome that matches the construct being studied, and (3) use multiple outcome measures whenever possible (Schraw, 2009).

This work has helped delineate how respondents interpret stems and scales for different constructs, however, the recommendations provided must align with the target construct of interest. If the priority for an assessment developer is to promote metacognition, that construct must be embedded in the instrument through its intended use as a formative tool. An instrument that will be used to collect measurement data for quantitative analysis should be designed with unidimensionality, validity, and reliability in mind for each chosen feature.

This work revealed two major target constructs that could be studied using confidence tiers: engaging learners in metacognition and measuring their self-evaluation of conceptual knowledge. Responsibility falls to the assessment developer to select which target construct best fits their goal for using a confidence tier and to use the most appropriate features for that construct. Within the framing of reflective interviews, respondents identified how confidence tiers allow for an assessor to identify if respondents are guessing on an assessment or using some degree of knowledge combined with test-taking strategies, emphasizing the importance of properly framing the purpose of self-reflective measurement and tailoring scales to serve assessors and learners most appropriately.

To improve metacognitive calibration, it has been shown that repeated self-assessment alone is not enough to improve this skill (Hawker et al., 2016; Webb & Karatjas, 2018). Literature from DBER fields has recommended that self-assessment should be used in conjunction with metacognitive training, reviewing one's own past performance, and benchmarking, i.e. comparing one's performance against that of others (Pazicni & Bauer, 2014). With this recommendation in mind, assessors should consider using metacognitive prompting within formative settings to echo the 'easing in' process that has been recommended for decades. Formative metacognitive training allows assessment developers who use a confidence tier tool to conduct more accurate measurement and identification of alternate conceptions with a population who is more metacognitively experienced. Ultimately, these recommendations should allow for a) improved metacognitive prompting and b) valid and reliable data regarding formal reasoning and alternative conceptions to be collected to improve future instruction.

Limitations

This study was conducted at one institution with a narrow sample population who had completed an introductory STEM course. While the interview protocol leveraged stems and scales from previous instruments that spanned contexts, this study was conducted within the single context of a General Chemistry assessment and reflection on content knowledge was done within that context. The sample collected was limited in size to ten interviewees, though interviews were continued until saturation was reached across the major observable themes. Despite the constrained sample size, the findings which emerged from this rich qualitative data provide a strong foundation for the recommendations to tailor stems and scales which reflect respondents' interpretations of confidence tiers. These findings converge with previous qualitative work demonstrating that confidence tiers can be subjective and result in use of unreliable metrics such as test-taking strategies (Koevoets-Beach et al., 2023). In examples outside of education like professional certification exams, assessment designers are advised that novices and experts within narrow sample populations may interpret and respond to confidence tiers differently (Fraundorf et al., 2022). Given that our findings converge with existing work across contexts, this supports our findings being transferable to a study conducted with a larger sample size. One further consideration for the limited sample size is that

the linguistic or cognitive distinctions apparent to college students may not generalize to younger learners or English language learners. Further investigation is required to better understanding how their engagement may differ.

It may be noted that no demographic data was obtained for the interviewees or reported for this study. The influence of gender and intersectional identities has been observed for confidence tiers in quantitative studies (Lundeberg et al., 1992) but was considered outside of the scope of this study. The scope of language and design features used in confidence tiers up to this point provides enough breadth that learners may rely on variable external factors which can be highly influenced by their diverse lived experiences to make their confidence judgments. It is our hope that through implementation of these highly tailored recommendations for confidence tier stems and scales, equity implications can more aptly be addressed in future studies as the tool is more narrowly focused on metacognition and content knowledge.

This work focused directly on using learners' interpretations to generate recommendations which directly prompt self-reflection and provide meaningful measurement of the target constructs, however, we can acknowledge that the utility of confidence tiers is dependent on individuals' willingness to engage with them. Our recommendations seek to increase meaningful engagement with confidence tiers, but a single exposure with this type of metacognitive prompting is not enough to improve accuracy in self-assessment calibration. With repeated exposure to these types of measures and regular practice, confidence tiers can become more reliable tools for both the learner and the assessor.

The recommendations generated from literature review and qualitative interviews have been largely directed towards developers of research-based assessments, particularly in DBER fields. While we believe these recommendations include widely applicable features for classroom assessment, they do focus on threats to validity and reliability of data in novel instrument design which may be of less concern in non-research contexts. In contexts outside of the classroom, these considerations may still be valuable for assessment development which seeks to target self-regulated learning in specific participant populations. Extensive quantitative literature has been reported regarding the psychometric qualities of different scale types and these studies were considered to be part of the rationale for conducting this qualitative study. Future work comparing the psychometric qualities of an assessment which used both an original confidence tier and a revised version based on our recommendations would further solidify the findings from our cognitive interviews.

Conclusions

This study highlighted how design differences in two-tiered confidence items can have distinct impacts on respondents' interpretations and perceived utility. Examination of how learners understand and interpret these components allowed us to identify both strategies and challenges that affect their confidence judgments. Our analysis revealed deeper understanding of the misalignment that can occur when design is not rooted in learners' mental models, justifying the need to tailor confidence stems and scales to reflect the construct of interest. Our recommendations first require an assessment designer to identify whether the target for using a confidence tier is to improve/promote metacognition or to measure learners' self-evaluation of their confidence in assessed content.

If targeting metacognitive practice:

- Use should be in low-stakes, formative assessments
- Stems should use specific language targeting "strength of understanding"

- Scales may be odd-numbered but with the midpoint option labeled as "made an informed guess" or "felt unsure" to direct reflection
- Scales should be constrained between 4 and 7 options
- Multiple interactions with the tier should be embedded in the design, rather than a single use

If targeting measurement of self-evaluation of conceptual knowledge:

- Use can be in either formative or summative assessments
- Stems should be narrowly tailored to prompt reflection on the exact context of the assessment
- Scales are recommended to be even numbered to provide more analytical power and eliminate the "dumping ground" of "average"
- Scales should provide 4-6 descriptions with paired numerical values to direct reflection which can be analyzed as interval data
- Scale descriptions should target uncertainty and informed guessing to provide actionable interpretations for assessors

These evidence-based recommendations are intended to improve the clarity and utility of confidence tiers, promote a more meaningful evaluation of understanding from both learners and assessors, and strengthen the quality of data used to make claims about respondents' cognition and/or metacognition. We intend to lay foundation for further research which investigates the potential effects of assessment design on learners' perceptions and understanding. By identifying and attending to those effects, assessors can develop more robust confidence claims and target learners who may benefit from more direction in developing their skills to engage with self-regulated learning.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Acknowledgements

We would like to acknowledge Karen Julian for her help with the early stages of this project and the other members of the Balabanoff group for their insight. We would also like to thank our student participants for their thoughtful engagement and responses during their interviews.

Received: 2/21/2025. **Accepted:** 10/16/2025. **Published:** 11/20/2025.

Citation: Koevoets-Beach, C., Kurdi, D., & Balabanoff, M. (2025). Considerations for designing measures of confidence. *Practical Assessment, Research, & Evaluation*, 30(1)(9). Available online: https://doi.org/10.7275/pare.2919

Corresponding Author: Morgan Balabanoff, University of Lousiville.
Email: morgan.balabanoff@louisville.edu

References

Abell, T. N., & Bretz, S. L. (2019). Development of the Enthalpy and Entropy in Dissolution and Precipitation Inventory. *Journal of Chemical Education*, *96*(9), 1804–1812. https://doi.org/10.1021/acs.jchemed.9b00186

Adey, P., Shayer, M., & Yates, C. (1995). *Thinking science: The curriculum materials of the Cognitive Acceleration through Science Education (CASE) project* (2nd ed.). Nelson. https://cir.nii.ac.jp/crid/1130000794564041344

Aitken, R. C. (1969). Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine*, *62*, 89–93.

Ajzen, I., & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology*, *22*(5), 453–474. https://doi.org/10.1016/0022-1031(86)90045-4

Albaum, G. (1997). The Likert Scale Revisited. *Market Research Society*, *39*(2). https://doi.org/10.1177/147078539703900202

Al-Rubayea, A. A. M. (1996). *An analysis of Saudi Arabian high school students' misconceptions about physics concepts* [Ph.D., Kansas State University]. https://www.proquest.com/docview/304309372/abstract/B5303F0B62934CA9PQ/1

Archer, N. S. (1962). A Comparison of the Conventional and Two Modified Procedures for Responding to Multiple-Choice Test Items. *The Yearbook of the National Council on Measurement in Education*, *19*, 78–82.

Assimi, E., Janati Idrissi, R., Zerhane, R., & Boubih, S. (2024). The use of a three-tier diagnostic test to investigate conceptions related to cell biology concepts among pre-service teachers of life and earth sciences. *Journal of Biological Education*, *58*(4), 864–891. https://doi.org/10.1080/00219266.2022.2134175

Balabanoff, M., Fulaiti, H. A., DeKorver, B., Mack, M., & Moon, A. (2022). Development of the Water Instrument: A comprehensive measure of students' knowledge of fundamental concepts in general chemistry. *Chemistry Education Research and Practice*, *23*(2), 348–360. https://doi.org/10.1039/D1RP00270H

Bandura, A. (1977). *Social Learning Theory*. Prentice-Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Bell, P., & Volckmann, D. (2011). Knowledge Surveys in General Chemistry: Confidence, Overconfidence, and Performance. *Journal of Chemical Education*, *88*(11), 1469–1476. https://doi.org/10.1021/ed100328c

Bergman, R. D. (2009). *Testing the measurement invariance of the Likert and graphic rating scales under two conditions of scale numeric presentation* [Ph.D., The University of Nebraska - Lincoln]. https://www.proquest.com/docview/304940435/abstract/7A2DE5CA716D40A2PQ/1

Blank, L. M. (2000). A metacognitive learning cycle: A better warranty for student understanding? *Science Education*, *84*(4), 486–506. https://doi.org/10.1002/1098-237X(200007)84:4<486::AID-SCE4>3.0.CO;2-U

Borman, G. D., & Overman, L. T. (2004). Academic Resilience in Mathematics among Poor and Minority Students. *The Elementary School Journal*, *104*(3), 177–195. https://doi.org/10.1086/499748

Brandriet, A. R., & Bretz, S. L. (2014). Measuring meta-ignorance through the lens of confidence: Examining students' redox misconceptions about oxidation numbers, charge, and electron transfer. *Chemistry Education Research and Practice*, *15*(4), 729–746. https://doi.org/10.1039/C4RP00129J

Bunce, D. M., Schroeder, M. J., Luning Prak, D. J., Teichert, M. A., Dillner, D. K., McDonnell, L. R., Midgette, D. P., & Komperda, R. (2023). Impact of Clicker and Confidence Questions on the Metacognition and Performance of Students of Different Achievement Groups in General Chemistry. *Journal of Chemical Education*, *100*(5), 1751–1762. https://doi.org/10.1021/acs.jchemed.2c00928

Burney, G. M. (1974). *The Construction and Validation of an Objective Formal Reasoning Instrument* [Doctor of Education dissertation]. Univeristy of Northern Colorado.

Caleon, I. S., & Subramaniam, R. (2010). Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions. *Research in Science Education*, *40*(3), 313–337. https://doi.org/10.1007/s11165-009-9122-4

Caleon, I., & Subramaniam, R. (2010). Development and Application of a Three-Tier Diagnostic Test to Assess Secondary Students' Understanding of Waves. *International Journal of Science Education*. https://www.tandfonline.com/doi/full/10.1080/09500690902890130?needAccess=true

Casselman, B. L., & Atwood, C. H. (2017). Improving General Chemistry Course Performance through Online Homework-Based Metacognitive Training. *Journal of Chemical Education*, *94*(12), 1811–1821. https://doi.org/10.1021/acs.jchemed.7b00298

Chyung, S. Y. (Yonnie), Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement*, *56*(10), 15–23. https://doi.org/10.1002/pfi.21727

Connor, M. C., Glass, B. H., & Shultz, G. V. (2021). Development of the NMR Lexical Representational Competence (NMR-LRC) Instrument As a Formative Assessment of Lexical Ability in 1H NMR Spectroscopy. *Journal of Chemical Education*, *98*(9), 2786–2798. https://doi.org/10.1021/acs.jchemed.1c00332

Cook, E., Kennedy, E., & McGuire, S. Y. (2013). Effect of Teaching Metacognitive Learning Strategies on Performance in General Chemistry Courses. *Journal of Chemical Education*, *90*(8), 961–967. https://doi.org/10.1021/ed300686h

Coombs, C. H. (1953). On the Use of Objective Examinations. *Educational and Psychological Measurement*, *13*(2), 308–310. https://doi.org/10.1177/001316445301300214

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The Assessment of Partial Knowledge. *Educational and Psychological Measurement*, *16*(1), 13–37. https://doi.org/10.1177/001316445601600102

Davis, E. A. (1996). *Metacognitive Scaffolding To Foster Scientific Explanations*. https://eric.ed.gov/?id=ED394853

de Finetti, B. (1965). Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item. *British Journal of Mathematical and Statistical Psychology*, *18*(1), 87–123. https://doi.org/10.1111/j.2044-8317.1965.tb00695.x

Dori, Y. J., Avargil, S., Kohen, Z., & Saar, L. (2018). Context-based learning and metacognitive prompts for enhancing scientific text comprehension. *International Journal of Science Education*, *40*(10), 1198–1220. https://doi.org/10.1080/09500693.2018.1470351

Ebel, R. L. (1965). Confidence Weighting and Test Reliability. *Journal of Educational Measurement*, *2*(1), 49–57.

Echternacht, G. J. (1972). The Use of Confidence Testing In Objective Tests. *Review of Educational Research*, *42*(2), 217–236. https://doi.org/10.3102/00346543042002217

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Flynn, D., van Schaik, P., & van Wersch, A. (2004). A Comparison of Multi-Item Likert and Visual Analogue Scales for the Assessment of Transactionally Defined Coping Function1. *European Journal of Psychological Assessment*, *20*(1), 49–58. https://doi.org/10.1027/1015-5759.20.1.49

Franklin, B. J. (1992). *The development, validation, and application of a two-tier diagnostic instrument to detect misconceptions in the areas of force, heat, light and electricity* [Ph.D., Louisiana State University and Agricultural & Mechanical College]. https://www.proquest.com/docview/304008535/abstract/911D6E63C6964051PQ/1

Fraundorf, S., Caddick, Z., Rottman, B., Nokes-Malach, T., Swanson, D., Bazemore, A., O'Neill, T., & Lipner, R. (2022). *Conceptual Foundations for Designing Continuing Certification Assessments for Physicians* [White paper]. American Board of Medical Specialties. https://www.abms.org/wp-content/uploads/2022/07/conceptual-foundations-continuing-certification-assessments-for-physicians.pdf

Freeman, J. G., Stoch, S. A., Chan, J. S. N., & Hutchinson, N. L. (2004). Academic Resilience: A Retrospective Study of Adults With Learning Difficulties. *Alberta Journal of Educational Research*, *50*(1), Article 1. https://doi.org/10.11575/ajer.v50i1.55038

Garland, R. (1991). The Mid-Point on a Rating Scale: Is it Desirable? *Marketing Bulletin*, *2*(1), 66–70.

Garner, R., & Alexander, P. A. (1989). Metacognition: Answered and Unanswered Questions. *Educational Psychologist*, *24*(2), 143–158. https://doi.org/10.1207/s15326985ep2402_2

Georghiades, P. (2000). Beyond conceptual change learning in science education: Focusing on transfer, durability and metacognition. *Educational Research*, *42*(2), 119–139. https://doi.org/10.1080/001318800363773

Georghiades, P. (2004). From the general to the situated: Three decades of metacognition. *International Journal of Science Education*, *26*(3), 365–383. https://doi.org/10.1080/0950069032000119401

Glaser, B. G. (1965). The Constant Comparative Method of Qualitative Analysis. *Social Problems*, *12*(4), 436–445. https://doi.org/10.2307/798843

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93–103. https://doi.org/10.3758/MC.36.1.93

Händel, M., & Fritzsche, E. S. (2015). Students' confidence in their performance judgements: A comparison of different response scales. *Educational Psychology*, *35*(3), 377–395. https://doi.org/10.1080/01443410.2014.895295

Hansen, R. (1971). The Influence of Variables Other than Knowledge on Probabilistic Tests. *Journal of Educational Measurement*, *8*(1), 9–14.

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the Certainty of Response Index (CRI). *Physics Education*, *34*(5), 294–299.

Hawker, M. J., Dysleski, L., & Rickey, D. (2016). Investigating General Chemistry Students' Metacognitive Monitoring of Their Exam Performance by Measuring Postdiction Accuracies over Time. *Journal of Chemical Education*, *93*(5), 832–840. https://doi.org/10.1021/acs.jchemed.5b00705

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*, 98–99.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186–201. https://doi.org/10.1037/h0074579

Hernandez, P. R., Hopkins, P. D., Masters, K., Holland, L., Mei, B. M., Richards-Babb, M., Quedado, K., & Shook, N. J. (2018). Student integration into STEM careers and culture: A longitudinal examination of summer faculty mentors and project ownership. *CBE—Life Sciences Education*, *17*(3), ar50. https://doi.org/10.1187/cbe.18-02-0022

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. *Physics Teacher*, *30*(3), 141–158.

Hevner, K. (1932). *A Method of Correcting for Guessing in True-False Tests and Empirical Evidence in Support of It* (world). https://www.tandfonline.com/doi/abs/10.1080/00224545.1932.9919159

Hilbert, S., Kuchenhoff, H., Sarubin, N., Nakagawa, T. T., & Buhner, M. (2015). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a Likert-type, and a visual analogue scale. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *1*, 3–24. https://doi.org/10.4473/TPM23.1.1

Hill, G. D. (1997). *Conceptual change through the use of student-generated analogies of photosynthesis and respiration by college non-science majors* [Ed.D., University of Georgia]. https://www.proquest.com/docview/304355289/abstract/671A5B2DB7484F74PQ/1

Hollingworth, H. L. (1913). *Experimental Studies in Judgment*. Science Press.

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital*, *4*(1), 100–113. https://doi.org/10.1108/14691930310455414

Jacobs, J. E., & Paris, S. G. (1987). Children's Metacognition About Reading: Issues in Definition, Measurement, and Instruction. *Educational Psychologist*, *22*(3–4), 255–278. https://doi.org/10.1080/00461520.1987.9653052

Jacobs, S. S. (1971). Correlates of Unwarranted Confidence in Responses to Objective Test Items. *Journal of Educational Measurement*, *8*(1), 15–20. https://doi.org/10.1111/j.1745-3984.1971.tb00901.x

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y

Koevoets-Beach, C., Julian, K., & Balabanoff, M. (2023). " I guess it was more than just my general knowledge of chemistry ": Exploring students' confidence judgments in two-tiered assessments. *Chemistry Education Research and Practice*, *24*(4), 1243–1261. https://doi.org/10.1039/D3RP00127J

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle Response Functioning in Likert-responses to Personality Items. *Journal of Business and Psychology*, *22*(3), 251–259. https://doi.org/10.1007/s10869-008-9064-2

Lavi, R., Shwartz, G., & Dori, Y. J. (2019). Metacognition in Chemistry Education: A Literature Review. *Israel Journal of Chemistry*, *59*(6–7), 583–597. https://doi.org/10.1002/ijch.201800087

Lawson, A. (1978). Relationships among Performances on Group Administered Items of Formal Reasoning. *Perceptual and Motor Skills*, *48*(1), 71–78. https://doi.org/10.2466/pms.1979.48.1.71

Lee, J., & Paek, I. (2014). In Search of the Optimal Number of Response Categories in a Rating Scale. *Journal of Psychoeducational Assessment*, *32*(7), 663–673. https://doi.org/10.1177/0734282914522200

Liampa, V., Malandrakis, G. N., Papadopoulou, P., & Pnevmatikos, D. (2019). Development and Evaluation of a Three-Tier Diagnostic Test to Assess Undergraduate Primary Teachers' Understanding of Ecological Footprint. *Research in Science Education*, *49*(3), 711–736. https://doi.org/10.1007/s11165-017-9643-1

Lozano, L. M., Garcia-Cueto, E., & Muniz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology*, *4*(2). https://econtent.hogrefe.com/doi/abs/10.1027/1614-2241.4.2.73

Lundeberg, M. A., Fox, P. W., & LeCount, J. (1992, April 21). *Highly Confident, but Wrong: Gender Differences and Similarities in Confidence Judgments*. Annual Meeting of the American Educational Research Association, San Francisco (CA).

Manstead, A. S. R., & van Eekelen, S. A. M. (1998). Distinguishing Between Perceived Behavioral Control and Self-Efficacy in the Domain of Academic Achievement Intentions and Behaviors. *Journal of Applied Social Psychology*, *28*(15), 1375–1392. https://doi.org/10.1111/j.1559-1816.1998.tb01682.x

Matell, M., & Jacoby, J. (1971). Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity. *Educational and Psychological Measurement*, *31*(3), 657–674. https://doi.org/10.1177/001316447103100307

McClary, L. M., & Bretz, S. L. (2012). Development and Assessment of A Diagnostic Tool to Identify Organic Chemistry Students' Alternative Conceptions Related to Acid Strength. *International Journal of Science Education*, *34*(15), 2317–2341. https://doi.org/10.1080/09500693.2012.684433

McDonald, R. P. (2013). *Test Theory: A Unified Treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications, Inc.

Miller, M. (2002). Resilience elements in students with learning disabilities. *Journal of Clinical Psychology*, *58*(3), 291–298. https://doi.org/10.1002/jclp.10018

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive Monitoring Accuracy and Student Performance in the Postsecondary Classroom. *The Journal of Experimental Education*, *74*(1), 7–28.

Ning, H. K., & Downing, K. (2015). A latent profile analysis of university students' self-regulated learning strategies. *Studies in Higher Education*, *40*(7), 1328–1346. https://doi.org/10.1080/03075079.2014.880832

Pazicni, S., & Bauer, C. F. (2014). Characterizing illusions of competence in introductory chemistry students. *Chem. Educ. Res. Pract.*, *15*(1), 24–34. https://doi.org/10.1039/C3RP00106G

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology*, *77*(3), 336–342. https://doi.org/10.1037/0021-9010.77.3.336

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Soderquist, H. O. (1936). A new Method of Weighting Scores in a True-False Test. *The Journal of Educational Research*. https://www.tandfonline.com/doi/abs/10.1080/00220671.1936.10880670

Sparks, P., Guthrie, C. A., & Shepherd, R. (1997). The Dimensional Structure of the Perceived Behavioral Control Construct. *Journal of Applied Social Psychology*, *27*(5), 418–438. https://doi.org/10.1111/j.1559-1816.1997.tb00639.x

Staver, J. R., & Gabel, D. L. (1979). The Development and Construct Validation of a Group-Administered Test of Formal Thought. *Journal of Research in Science Teaching*, *16*(6), 535–544.

Sun, H., Zhou, Y., Culley, D. J., Lien, C. A., Harman, A. E., & Warner, D. O. (2016). Association between Participation in an Intensive Longitudinal Assessment Program and Performance on a Cognitive Examination in the Maintenance of Certification in Anesthesiology Program®. *Anesthesiology*, *125*(5), 1046–1055. https://doi.org/10.1097/ALN.0000000000001301

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology*, *82*(2), 306–314. https://doi.org/10.1037/0022-0663.82.2.306

Swineford, F. (1938). The measurement of a personality trait. *Journal of Educational Psychology*, *29*(4), 295–300. https://doi.org/10.1037/h0058735

Taslidere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. *Research in Science & Technological Education*, *34*(2), 164–186. https://doi.org/10.1080/02635143.2015.1124409

Terry, D. J., & O'Leary, J. E. (1995). The theory of planned behaviour: The effects of perceived behavioural control and self-efficacy. *British Journal of Social Psychology*, *34*(2), 199–220. https://doi.org/10.1111/j.2044-8309.1995.tb01058.x

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*(2), 264–273. https://doi.org/10.1348/135910710X510494

Trafimow, D., Sheeran, P., Conner, M., & Finlay, K. A. (2002). Evidence that perceived behavioural control is a multidimensional construct: Perceived control and perceived difficulty. *British Journal of Social Psychology*, *41*(1), 101–121. https://doi.org/10.1348/014466602165081

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Trow, W. C. (1923). The psychology of confidence. *Archives of Psychology*, *67*, 47–47.

Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The Effects of Feedback Timing on Learning Facts: The Role of Response Confidence. *Contemporary Educational Psychology*, *19*(3), 251–265. https://doi.org/10.1006/ceps.1994.1020

White, K. M., Terry, D. J., & Hogg, M. A. (1994). Safer Sex Behavior: The Role of Attitudes, Norms, and Control Factors. *Journal of Applied Social Psychology*, *24*(24), 2164–2192. https://doi.org/10.1111/j.1559-1816.1994.tb02378.x

Wiley, L. N., & Trimble, O. C. (1936). The ordinary objective test as a possible criterion of certain personality traits. *School and Society*, *43*, 446–448.

Yang, D.-C., & Sianturi, I. A. J. (2019). Assessing students' conceptual understanding using an online three-tier diagnostic test. *Journal of Computer Assisted Learning*, *35*(5), 678–689. https://doi.org/10.1111/jcal.12368

Yates, J. F. (1990). *Judgment and decision making* (pp. xvi, 430). Prentice-Hall, Inc.

Yip, M. C. W. (2007). Differences in Learning and Study Strategies between High and Low Achieving University Students: A Hong Kong study. *Educational Psychology*, *27*(5), 597–606. https://doi.org/10.1080/01443410701309126

Ziller, R. C. (1957). A measure of the gambling response-set in objective tests. *Psychometrika*, *22*(3), 289–292. https://doi.org/10.1007/BF02289129

Zimmerman, B. J. (1989). A Social Cognitive View of Self-Regulated Academic Learning. *Journal of Educational Psychology*, *81*(3), 329–339.

Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, *25*(1), 82–91. https://doi.org/10.1006/ceps.1999.1016

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, *53*(1), 141–160.