

A Mixed Methods Investigation into the Functionality of the Willingness to Consider
Contradictory Evidence Scale

Anna Zilberberg

Dena A. Pastor

James Madison University

Correspondence concerning this manuscript should be addressed to:

Anna Zilberberg
821 S. Main Street
MSC 6806
24 Anthony-Seeger Hall
Harrisonburg, VA 22807 USA
E-mail: zilberax@jmu.edu
Phone: 617-763-0349
Fax: 540-568-7878

Presented at the Annual Meeting of the Northeastern Educational Research Association, Rocky Hill, CT

A Mixed Methods Investigation into the Functionality of the Willingness to Consider Contradictory Evidence Scale

College educators are charged not only with imparting knowledge, but with developing students into better thinkers. As noted by Williams, Wise and West (2001), “although students coming into college already know how to think, one of the primary purposes of the college experience is to get them to think even better” (pg. 3). Improvement in thinking is often accomplished by teaching students critical thinking skills. Although what specific skills constitute critical thinking is debatable, researchers agree that the ability to consider a variety of different viewpoints and to incorporate these viewpoints in decision-making is essential to critical thought¹ (Baron, 2008; Peterson, 2004). The ability to consider another’s viewpoint is vital to a college education as it allows students to incorporate into their own understanding of the world the variety of different theories, perspectives, and opinions they will be exposed to throughout their college career and beyond.

In order to evaluate this cognitive disposition in students and draw conclusions about program effectiveness, educators need a viable means for assessing this trait. In fact, cognitive psychologists have developed numerous measures tapping into the different aspects of disposition toward critical thinking. However, the suitability of these measures for use in educational settings warrants careful attention. In the sections that follow, we describe different approaches that have been used to measure aspects of critical thinking and call attention to some of their weaknesses. We proceed to argue that it may be more effective and efficient to measure

¹ Various concepts in the critical thinking literature encompass this disposition. For example, cognitive scientists use terms such as “actively open-minded thinking”, “flexible thinking disposition”, “bracketing”, and “actively fair-minded thinking”. Interested readers should consult Petersen (2004) for a comprehensive review.

the extent to which a person *values* the consideration of contradictory viewpoints as opposed to trying to measure whether a person considers contradictory viewpoints themselves. We contend that Likert items could be an effective and efficient means by which to measure how much one values the consideration of contradictory viewpoints and identify items from a pre-existing instrument that could be used for such a purpose. We conclude the section by addressing the need for validity evidence to be collected for these items prior to their operational use.

Approaches to Measuring One's Ability to Consider Contradictory Viewpoints

A popular approach to measuring whether a respondent considers contradictory viewpoints involves asking the respondent to provide their opinion or view on an issue (e.g., Harvey, 1964; Suedfeld & Tetlock, 1977). Trained raters code these responses according to whether only one perspective was provided or whether differing perspectives on the issue were acknowledged. Although it may seem straightforward, this approach has many complications. First, hiring and training raters can be costly and time-consuming. Second, factors such as general intelligence, language fluency, and writing skills are likely to confound the ratings. Raters may be influenced by the mere eloquence of the verbal response or the nature of the favored opinions provided by the respondents (Peterson, 2004). Third, persuasive arguments are content-specific (i.e. a certain issue is the centerpiece of a response) and thus reflect the person's stance on a particular issue, and not the thinking used to arrive at that stance. Finally, only a small number of participants can be assessed via such measures, again due to the labor-intensive nature of rating the responses.

Given the complications in trying to measure the kind of thinking students do, it may be easier to measure the kind of thinking students value or what kind of thinking students consider to be good thinking. For example, Baron (1989) asked subjects how college students should

respond when encountering new information on such topics as politics or religion. He classified subjects according to whether or not they believed students should consider new information and if justified, possibly change their own views. This is a very straightforward approach to measuring the extent to which students value the consideration of contradictory views and opinions. Although raters are still required, it is easier to rate responses objectively, with resultant scores less influenced by irrelevant factors.

The problem with this approach is that we are measuring the type of thinking students value, not the type of thinking students do. That is, if a student reports that they believe a person should consider contradictory evidence, can we infer that the student themselves considers contradictory evidence? Not necessarily. Nonetheless, research indicates that valuing this certain thinking disposition and engaging in it are closely related. For instance, Baron (1989) found that subjects who valued the consideration of contradictory views were also more likely to recognize two-sided arguments provided by others as “good thinking” and provide differing perspectives on issues themselves.

A drawback in using Baron’s approach to measure the kind of thinking students value is the fact that raters are still needed. An alternative approach that does not require the use of raters is provided by Stanovich and West (1997). In their approach respondents were asked to use a Likert scale to indicate the extent to which they agreed with statements like “People should always take into consideration evidence that goes against their beliefs.” This statement is one of the 56 statements included on their Actively Open-Minded Thinking (AOT) scale². Because the AOT was created to measure thinking dispositions very broadly, statements assessing a wide

² The creation of the AOT to measure thinking dispositions was not the primary purpose of the research conducted by Stanovich and West (1997). Instead, Stanovich and West (1997) were interested in developing a method to assess myside bias, which is the extent to which an individual’s own views of opinions on an issue complicate their ability to evaluate the quality of the thinking of another on the issue. To this end they created the Argument Evaluation Test (AET), which requires the respondent to evaluate another person’s quality of thinking.

variety of different aspects of thinking were included. For example, in addition to items addressing one's willingness to consider contradictory evidence, items were also included to measure reflectivity, tolerance for ambiguity, absolutism, dogmatism, categorical thinking, superstitious thinking, and counterfactual thinking. Sá, West, and Stanovich (1999) also used the AOT in their research and added to the scale a belief identification subscale. Several of the items on this subscale along with one from the original AOT appear to be addressing the specific aspect of critical thought of interest here, which is one's willingness to consider contradictory evidence. Table 1 contains these items along with their corresponding AOT subscale and source.

There are many advantages to using the items in Table 1 to assess the extent to which students value the consideration of contradictory views and opinions. In comparison to some of the aforementioned approaches, this approach is advantageous in that no raters are needed to score the responses and the items can be scored objectively. After reverse scoring items 1, 2, and 3, the items are simply summed with the resulting score representing one's willingness to consider contradictory evidence. Because completion and scoring of the items requires little time and effort, a larger number of responses can be collected, thus increasing the power associated with any inferential tests involving the scores. As aforementioned, a disadvantage to this approach is that it is not measuring the type of thinking one does, but the type of thinking one values. However, given the complexities involved in accurately measuring the type of thinking one does, focusing on the type of thinking one values might be worthwhile, particularly since research has supported an association between the two.

Simply asking respondents to endorse the kinds of thinking styles they value seems to be a very straightforward and efficient means by which to obtain a glimpse into the thinking they engage in themselves when confronted with a differing perspective. Little research, however,

exists examining the extent to which this approach yields valid and useful information. To our knowledge, the only authors that have utilized these items in their research are Sá et al. and Stanovich and West. The items were not treated as a separate subscale, but were instead used and investigated as part of a larger scale, the AOT. Prior to using the sum of these items to represent one's openness to contradictory views, evidence is needed supporting the validity of the items in Table 1, which we have called the Willingness to Consider Contradictory Evidence Scale (WCCES). In the section that follows, we provide an overview of validity and the benefits of using a mixed-methods approach when gathering validity evidence for a scale.

What is Validity?

According to the *Standards for Educational and Psychological Testing* (1999), validity is “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p.9). Gathering validity evidence is therefore paramount if one wishes to make meaningful inferences from a score. In this case, we attempted to collect evidence supporting the notion that the items of the WCCES can be summed up to yield a total score indicating the extent to which students value consideration of contradictory evidence. Although there are many different approaches to collecting validity evidence for the measure, we focused on the structural validity and response process validity.

The structural validity pertains to the dimensionality of the scale. If items on the WCCES relate to each other in a predicted manner and tap into a single construct, then this finding would be considered evidence of structural validity for the WCCES. Gathering structural validity evidence requires the use of quantitative techniques (such as factor analysis). One of the first steps towards providing validity evidence for a scale is establishing that a total score reflects a

single characteristic. Thus, our first goal was to explore the dimensionality and internal structure of the WCCES.

The response process validity relates to the idea that the items actually elicit the response processes expected by researchers. This type of validity evidence provides information about why individuals are responding to the items they way they do. In the case of the WCCES, if an individual endorses a statement such as “beliefs should always be revised in response to new evidence”, then her response process should indicate that she values this type of thinking. Such tools as think-aloud interviews require participants to verbalize all thoughts as they are occurring while reading the test items and coming up with a response (Ericsson & Simon, 1993). Having respondents verbalize their thought processes as they are reading the WCCES items can help us figure out if the items and response scale are being interpreted by respondents in the manner intended.

In an attempt to gather comprehensive validity evidence for the WCCES, we decided to collect two types of validity evidence outlined above. To this end, we used a mixed methods approach. In the section that follows, we explain the rationale of the current study.

Purpose of the Current Study

The purpose of this study is to examine the functionality (i.e. gather validity evidence) of the WCCES, which measures the extent to which students’ value consideration of opposing viewpoints. To this end, a mixed methods sequential explanatory design was used. Such a design unfolded in two phases: (1) quantitative (exploratory factor analysis) and (2) qualitative (think-aloud interviews) (Creswell & Plano Clark, 2007). By combining quantitative and qualitative techniques in a sequential manner, we were able to explore the dimensionality of the scale and further probe the thinking processes elicited by the WCCES items. In this manner, the qualitative

analysis was used to expand and explain the quantitative analysis. This design allows researchers to interpret qualitative data in light of the statistical results found in the quantitative stage, and thereby gains a diversified understanding of the WCCES.

The current study unfolded in two phases. The first phase, Study A, used an exploratory factor analytic technique to investigate the dimensionality of the WCCES. The second phase, Study B, involved conducting think-alouds to explain the results from the Study A. The following sections describe these two studies in a consecutive order.

Study A

Method

Subjects and Procedures

The sample consisted of 1,014 incoming freshmen from a mid-sized, Southeastern university who were required to participate in a semi-annual institution-wide Assessment Day in August of 2008. The WCCES was part of a battery of instruments administered to students during a 3-hour testing session. Students were asked to mark the extent to which they agreed with each statement using a 6 point scale ranging from 1 (strongly disagree) to 6 (strongly agree). Students with out-of-range data or missing responses on the WCCES were eliminated from the sample, resulting in a final sample of 1,001 students. Demographic information for the sample is based only on the 988 students for which such information was available. Of these students, 64% were female, 81% were Anglo-American and the average age was 18.5 ($SD = 0.51$).

Analysis

Maximum likelihood exploratory factor analysis with oblique geomin rotation was used with solutions specifying 1 and 2 factors. All models were estimated using Mplus, version 5.2

(Muthén & Muthén, 2007). Three fit indices were used to decide upon a factor solution: the comparative fit index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the standardized root mean square residual (SRMR). Hu and Bentler (1998; 1999) recommend the use of CFI values greater than or equal to .95 and RMSEA and SRMR values at or below .06 and .08, respectively, in determining model fit. The chi-square statistic and likelihood ratio tests were not utilized since both are overly sensitive to sample size.

A parallel analysis was also used to determine which factor solution to retain. A parallel analysis entails an eigenvalue decomposition of an unreduced correlation matrix created from randomly generated data with characteristics similar to those of the original data (e.g., sample size, response scale). The eigenvalues based on the randomly generated data are plotted against those based on the sample data, with the factor solution just prior to the eigenvalues' point of intersection being favored.

Solutions favored by the three fit indices and the parallel analysis were further examined to identify the most interpretable solution. Both pattern and structure coefficients were studied, with values of $> |.40|$ for pattern coefficients used to designate an item as loading on a factor.

Results

Descriptive statistics for the items are shown in Table 1. The fact that the item means after reverse scoring are quite different from one another indicates problems with the scale. To illustrate, consider the situation where students, on average, value being open to contradictory evidence. In this scenario, we would anticipate that after reverse scoring, the means for all items would be in the agreement range of the response scale, which is above 3.5. If the opposite occurred and students, on average, did not value being open to contradictory evidence, we would anticipate the means for all items after reverse scoring to be below 3.5. As can be seen from

Table 1, the items differ from each other in their means after reverse scoring. The means of the first two items indicate that students, on average, are not open to contradictory evidence, whereas the means of the last three items indicate that students are open to contradictory evidence. The fact that the items are not telling us the same story about the average student is troubling.

Also troubling are the low correlations among the items, shown for items after reverse scoring in Table 2. Although all correlations are positive, they are all low in magnitude. The largest amount of variance shared between any two items is only 10%. These correlations indicate that responses to items are not strongly associated with each other; that is, responses to any one item are not strongly associated with responses to any other item.

We hesitantly proceeded with the factor analysis given the suggestion by some authors that factor analysis is not appropriate when a substantial number of correlations are below an absolute value of 0.30 (Hair, Anderson, Tatham & Black, 1998). The parallel analysis results, which are shown in Figure 1, favor a 1-factor solution over a 2-factor solution. The 1-factor solution was also considered more appropriate given the failure of the 2-factor solution to converge. Although a 1-factor solution was favored over a 2-factor solution, there was not strong support for the 1-factor model. Two of the three fit indices for the 1-factor solution indicated that the model did not fit the data ($CFI = 0.849$, $RMSEA = 0.108$, $SRMR = 0.046$). As well, the proportion of each item's variance accounted for by the factor was quite low. The pattern coefficients for items 1,2,3,4, and 5 equaled 0.52, 0.55, 0.39, 0.30, and 0.49, respectively. The square of a pattern coefficient represents the proportion of variance in the item explained by the factor. The largest percentage of an item's variance accounted for by the factor was 30%, which is far below the recommendation that the factor account for at least 50% of each item's variance

(Hair et al., 1995). Overall, these results indicate that the WCCES items do not tap into a single construct and also do not strongly relate to one other.

Study B

Method

The second phase of this mixed-methods study involved collecting and analyzing qualitative data in order to explain and expand the results of the quantitative study. More specifically, the goal of this qualitative study is to gather response process validity evidence for the WCCES and further explore the reasons why items of this instrument do not relate to each other in a hypothesized manner. The next sections describe the specific methods used, participants, and the results.

Participants

Out of sixty students randomly selected out of a pool of the students who took the WCCES during university-wide semi-annual assessment day in August 2008, eight students agreed to participate in the study (seven females and 1 male). As usual, all participants provided their informed consent for participating in the study. They received one hour of research participation credit to satisfy the requirement of the introductory psychology course. All participants were Caucasian. The average age was 18.54, and all of them were freshmen in college. The participants represented a variety of majors, including nursing, studio art, biology, and communication sciences.

Methods and Materials

Each session was conducted one-on-one with the researcher, who also played a role of the interviewer. Upon arriving at the location, the interviewer explained the details of the study including the fact that the session would be audio-taped and all information collected would be

kept confidential. Each participant was reminded that if they chose to participate, they would fulfill one hour of their research participation requirement through taking part in this study. Each participant was also informed that they could withdraw from the study at any point without any consequences of any kind.

Participants' think-aloud responses were recorded using a voice-recording software Free Sound Recorder 7.3.1 and a laptop computer.

Think-Aloud Procedure

The think-aloud procedure was explained to each participant first. Specifically, the researcher asked the participant to verbalize their thoughts as they read and provided their responses to a series of items. Participants were asked not to dwell on their thoughts or interpret their feelings, but simply to report their emerging thoughts. To get participants acquainted and comfortable with the think-aloud process, the first eight items in the think-aloud protocol served as practice items. The practice items included in this phase were similar in content and format to the WCCES items. In fact, the practice items were also created by Stanovich and colleagues and are commonly given along with the WCCES items. During this practice phase, the interviewer ensured that each participant understood all of the directions and could articulate his/her thoughts. After completing the practice session, the participant was asked to proceed with the think-aloud process on the next ten items. Out of these ten items, only five comprise the WCCES, which is the main focus of the present study. The other items were included to simulate the environment in which students first took the WCCES on the Assessment Day.

Coding Procedure

Overall, participants appeared to be engaged in the think-aloud process and there was no reason to suspect inaccuracy in the students' responses. Upon completion of the data collection stage, audio files were transcribed and coded by the researcher. The following sections describe this coding process.

Following the recommendation of Attride-Stirling (2001), the following data analytic steps were taken: (a) developing codes, (b) coding the data (text segments from the verbal protocol), (c) re-confirming data codes, and (d) organizing the data according to the codes. Iterative development of a coding framework marked the beginning of the qualitative data analysis. We followed the survey appraisal recommendations by Willis and Lessler (1999). These recommendations allowed us to remain attentive to the features characterizing sound survey items. More specifically, we were on the lookout for the following potential pitfalls commonly discovered through cognitive testing of a survey:

1. Clarity: Grammatical structure of the item makes it hard to read and understand; the statement is double-barreled; contains contradictory words; can be interpreted in various ways; is vague.
2. Assumption: The item contains an implicit assumption of a constant behavior or attitude.
3. Sensitivity/bias: The wording is sensitive in nature or contains bias; prompts a socially acceptable response set.
4. Response categories: The range of response options is inadequate or insufficient; categories are overlapping, missing, or stated in an illogical order.

Although these guidelines helped us focus our investigation, they did not constrain us. That is, we tried to code every theme that emerged from the data. Such an integrated coding

approach was best suited for this study, given its explanatory nature. The resulting coding framework is presented in Table 3. After coding the data, we proceeded to analyzing it.

Thematic Network Analysis

We used thematic network analysis (Attride-Stirling, 2001) to explore the patterns and themes emerging from the data. This analytic approach allows one to condense qualitative data into a meaningful “web-like” network, illustrating the interplay among basic, organizing, and global themes. In most traditional qualitative research, basic and organizing themes serve as building blocks for the global themes. That is, emerging lower-order themes yield global themes which are overarching ideas encapsulating the essence of the phenomenon. However, in this explanatory design study, the global theme is known – it is the *Construct Invalidity* of the WCCES. The current qualitative analysis allowed us to further probe sources of error contributing to the *Construct Invalidity*. These sources of error are reflected in the basic and organizing themes derived from the data.

Construction of the thematic network followed these steps. First, text segments were organized by codes and re-analyzed to find *basic themes*, which are “specific enough to be discrete...and broad enough to encapsulate a set of ideas contained in numerous text segments” (Attride-Stirling, 2001, p. 392). Then, logical groupings of the basic themes led to the *organizing themes* which were interpreted to be contributing directly to the *global theme*. The resulting thematic network is presented in Figure 2, and described below.

Basic and Organizing Themes

Six basic themes were constructed from the codes. These basic themes were then organized into the following organizing themes: *Inadequate Range of Responses*, *Construct*

Irrelevancy, Socially Desirable Response Set, and Lack of Clarity. The sections that follow discuss these organizing themes as well as their contributing basic themes.

Lack of Clarity. The first organizing theme *Lack of Clarity* encapsulates the following basic themes: (a) *grammar/style*, (b) *multiple interpretations*, (c) *condition-dependent*. Certain participants' responses and behaviors led to the emergence of the first basic theme – *grammar/style*. For example, if a respondent re-read the item, this behavior was interpreted to indicate a poorly constructed and thus confusing sentence, and that item was flagged for *grammar/style*. Similarly, remarks such as “This question is confusing” and “I am not sure what this means” were also considered to be indicators of subpar grammar, style, or word use. All items but number 4 scored high in this category.

Multiple interpretations theme subsumed items that prompted different interpretations from different respondents. To illustrate, consider the following response provided by a student to Item 3: “They could go a lot of different ways, like your definition of ‘beliefs’, your definition of ‘consideration of evidence’”. In the same vein, another student pondered: “When you mean – consideration – do you mean – thinking about the evidence or incorporating it into your belief?” Clearly, there were multiple interpretations of phrases such as “consideration of evidence” and students' responses fluctuated due to the ambiguity of this term. Items 3 and 4 were especially likely to cause confusion in this regard. The word “belief” and its derivatives warrant special attention as this word was most likely to elicit very different interpretations. All of the respondents interpreted the word “belief” as a “religious belief”. Some equated the verb “believe” to the verb “hope”. Others thought of political beliefs. One respondent used “believe” in the sentence “I believe in my family,” suggesting that this word can be used to signify trust or

reliance. Items containing this word (all of them) were therefore likely to produce inconsistent responses.

The third basic theme contributing to the *Lack of Clarity* was labeled as *condition-dependent*. It pertains to the observation that students' opinions (and thus responses on the WCCES items) tended to change depending on what exactly they were thinking about. For instance, one student noted during the structured interview that "a lot of things are based on the circumstances" and that his opinions change "depending on what specifically he is thinking about". Similarly, another student noted that "loyalty depends on the circumstances". Yet another said that modifying a belief based on the evidence "depends on the evidence and depends on a belief". Items 3 and 4 ranked high in this category.

Socially Desirable Response Set. This organizing theme refers to the finding that certain items prompted student to give socially desirable responses. Social desirability refers to the test-takers' inclination to provide a socially acceptable, but not necessarily accurate, response (Willis & Lessler, 1999). The results of this qualitative investigation of the WCCES led us to believe that many items on this scale did elicit socially desirable responses. Two organizing themes made up this category: (1) *positive connotations* and (2) *negative connotations*. Specifically, item 1 contains a word with a very strong positive connotation – "persevere". Not surprisingly, students endorsed this item almost uniformly. The rationale provided by students during the think-alouds solidified our suspicion. For example, one student said that "persevering in one's beliefs is noble". More broadly, this thematic category applied to all of the items because the word "belief" also has a positive connotation. For instance, one participant said that "beliefs are a core of who someone is; it is a solid foundation of your personality". The second organizing theme *negative connotations* referred to the items worded in an unappealing way. Two items

were flagged in this category – item 2 and item 5. Item 2 has a word “disregard” and item 5 has a word “abandon”. A few participants noted that “disregarding evidence completely is just ignorance”. One student said that whenever she sees the word “abandon” she thinks about “abandoning children” (speak of negative associations!).

Construct Irrelevancy. The main organizing theme that contributes to this thematic category is *religion*. The topic of religion kept coming up during the think-alouds and throughout the structured interviews. Nearly every participant noted that “the word belief makes them think about religion” or a similar remark along these lines. When thinking of examples during the interviews, religious beliefs kept coming up as examples. One girl told a story of how her religious beliefs were tested with evidence, but how she persevered in her faith despite of that. It seems that religion keeps surfacing as an irrelevant construct in this instrument and contributes to the overall *Construct Invalidity* of the scale.

Inadequate Range of Responses. This theme is narrow enough that it does not need an organizing theme. Respondents indicated that the Likert scale response options did not provide a sufficient range of responses. They wanted to see an option that would cover “the middle ground”, such as “I don’t know”, “undecided”, or “neither agree nor disagree”. Though having these response options might deter test-takers from the extreme ratings (and thus jeopardize the psychometric analyses of the scale), such middle ground seems to be necessary nonetheless. By omitting this option, researchers do not give the respondents an opportunity to provide a fully accurate account of their attitudes.

Discussion

Our study focused on an essential aspect of critical thought and open-minded thinking, which is an individual's ability to consider viewpoints different than their own. Given the

complexities in measuring the extent to which one engages in this activity, we focused instead on assessing the extent to which one values this aspect of open-minded thinking and identified a subset of items from the AOT deemed suitable for this purpose.

Because the items of the WCCES had yet to be used in isolation to measure the extent to which one values this aspect of critical thinking, our study sought to collect validity evidence for the scale. We used a mixed methods approach to examine the functioning of the items, with Study A using the correlations between items and exploratory factor analyses to investigate the structural validity of the scale and Study B using think-alouds and structured interviews to explore if the items were functioning as intended. A summary of the results of Study A and Study B are summarized below followed by our recommendations for researchers interested in measuring this construct.

Study A

The results of Study A indicated that the responses to the items were not strongly associated with one another. Because respondents were not responding in the same way across items, we had little support for the claim that items were measuring the same characteristic. This conclusion was further supported by the results of the exploratory factor analyses. Although a 1-factor model was favored over 2-factor model, the 1-factor model failed to fit the data. As a whole, our quantitative analyses did not yield supporting evidence for the structural validity of the scale. Our results indicated that the sum of the items could not be meaningfully interpreted as reflecting a single characteristic given that respondents were providing differing responses to each the five WCCES items.

Study B

Although our quantitative study was incredibly useful in alerting us to problems with the scale, it did not provide an explanation for why individuals were responding inconsistently across items. To answer this question, we embarked on a qualitative study. Our think alouds provided not only several explanations as to why respondents were providing inconsistent responses across items, but they also pointed to problems with the scale as a whole. Below we discuss what we consider to be the most important findings from our qualitative study and their implications.

Use of the term “belief”. Perhaps the most important finding from our qualitative study was what we learned about the respondents’ interpretation of the word “belief”, which is present in all WCCES items. Many of the respondents considered the term “belief” to mean religious beliefs and used that context in responding to all items. Although a good critical thinker should value being open to differing religious beliefs, these items were not written to address religious beliefs specifically. The term “belief” was instead intended to be more general. There are two ways to interpret this finding, both indicating problems with the scale. The first considers this interpretation of the term belief as being a source of construct irrelevant variance in the responses. In other words, because respondents are thinking about religious beliefs and not beliefs in general, a factor that is irrelevant to the construct of interest is systematically influencing responses. This view considers preferences in how one thinks about religious beliefs to be irrelevant or not central to preferences in how one thinks about beliefs in general. If instead one considers preferences in how one thinks about religious beliefs to be relevant, then the limited interpretation respondents are using for the term “belief” may be contributing to construct underrepresentation. Although the measure is intended to be measuring beliefs more

broadly, the interpretation respondents are imposing on the term results in a more limited measure, one that addresses only religious beliefs.

This latter interpretation would be more plausible if all respondents were thinking about religious beliefs on every item, but this was not the case. Although religion was brought up repeatedly and by all respondents during Study B, it was not the interpretation utilized by every respondent on every item. Some respondents considered political beliefs and others questioned more generally what was meant by the term “belief”. Given the ambiguity associated with the term and the religious connotation respondents naturally impose upon it, we recommend that future researchers use another term (e.g., views, viewpoints, perspectives, opinions). If the term “belief” is retained, at the very least it should be defined for the respondents. For example, respondents could be told: “The term ‘belief’ in the following items refers to any belief. It could be a religious belief, political belief, or a belief you hold about yourself or another individual or group.”

Other wording effects. Terms other than “belief” used in the items were also identified as being problematic. For instance, “persevere” has a positive connotation and respondents were unwilling to claim that they would not persevere in something. Similarly, “abandon” has a negative connotation and respondents were unwilling to claim that they would ever abandon something. Thus, the positive or negative connotations associated with certain terms prompted socially desirable responses. These particular wording effects are sources of systematic error variance as they elicit predictable response behavior that is irrelevant to the construct. For this reason, future researchers should avoid terms like “persevere”, “abandon” and perhaps “disregard” in Likert items used to assess this construct.

It depends. Many respondents were uncertain as to how to respond to the items because they claimed their answer “depends on the evidence and depends on the belief”. In fact, many wished they had the response options of “neither agree nor disagree” or “I don’t know” or “undecided” to the items for this reason. We agree with the respondents and assert that answering “it depends” to many of the items on the WCCES may actually be indicative of a higher level of critical thinking than a response of “strongly disagree” to items 1, 2, and 3 and “strongly agree” to items 4 and 5. To illustrate this point, consider item 5 “Beliefs should always be revised in response to new information or evidence”. A response of “Strongly Agree” to this item is supposed to be desirable. However, if someone always revised their beliefs in response to new evidence, they would *not* be engaging in good critical thinking. Good critical thinking would instead be characterized by discerning the quality of the evidence and only altering one’s belief if the strength of the evidence warranted such an action. Thus, the answer of a good critical thinker to this item and the others that contain the word “evidence” would be “it depends”, not “Strongly Agree”.

Similarly, we would expect good critical thinkers to respond “Strongly Agree” or “it depends” as opposed to “Strongly Disagree” to item 3 “Certain beliefs are just too important to abandon no matter how good a case can be made against them”. Even the best critical thinker has beliefs they would not reconsider because no evidence could possibly exist that is strong enough to refute the belief (e.g., the belief that smoking causes cancer). There is no simple solution for overcoming these problems with the scale. Qualifying what is meant by evidence may help somewhat, so that respondents know that the “evidence” being referred to in the items is reasonable, relevant, strong and of high quality. Trying to overcome the problem with the term

“belief” may somewhat be alleviated by our previous recommendations, which include using a different term for belief or defining what is meant by the term.

Future Directions

The results of our quantitative and qualitative studies combined suggest that the WCCES should not be used to measure the extent to which a person values the consideration of views different from their own. Although the Likert items investigated in the current study cannot be used, we still believe that Likert items could be a useful tool for assessing this characteristic and hope that researchers will use our findings to guide future item development.

Implications for AOT

Because the items selected for the WCCES are a subset of items included on the 1999 version of the Actively Open-Minded Thinking (AOT) scale developed by Sá, West, and Stanovich, our findings have implications for that scale. The AOT consists of the WCCES items as well as items measuring reflectivity, tolerance for ambiguity, absolutism, dogmatism, categorical thinking, superstitious thinking, and counterfactual thinking. All items on the AOT are summed to create a total score, which is interpreted as measuring open-minded thinking. Our finding that the WCCES items could not be summed to create a meaningful score calls into question whether the AOT items, of which the WCCES is a part, can also be summed to create a meaningful score. Our study therefore reinforces the need for studies examining the structural validity of the AOT. This is particularly important since the problems identified with the WCCES items are likely to still occur when these items are given as part of the AOT.

Mixed methods in validity research

Using both quantitative and qualitative methods to examine the validity of the WCCES scores was incredibly informative. Although the quantitative part of our study allowed to assess

the dimensionality of our items and alerted us to their low intercorrelations, it did not answer the question of why the item correlations were so low.

Without the qualitative findings, we would have been left speculating as to why responses were so inconsistent across items and calling for further research. The qualitative part of our study therefore filled an important gap. The think-alouds provided us with rich information about how the items were functioning. Having both pieces of information, both the quantitative and the qualitative, strengthened the conclusions we made regarding the WCCES and allowed us to provide useful guidelines for future researchers wanting to measure the construct.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.
- Attride-Stirling, J. (2001). Thematic networks: An analytical tool for qualitative research. *Qualitative Research, 1*, 385-405.
- Baron, J. (2008). *Thinking and Deciding, 4th Ed.* New York: Cambridge University Press.
- Baron, J. (1989). Why a theory of social-intelligence needs a theory of character. In R. S. Wyer & T. K. Srull (Eds.), *Advances in social cognition, Vol. 2: Social intelligence and cognitive assessments of personality* (pp. 61-70). Hillsdale, NJ : Erlbaum.
- Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis*. Cambridge, MA: The MIT Press.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis with readings* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Harvey, O. J. (1964). Some cognitive determinants of influencibility. *Sociometry, 27*, 208-221.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Petersen, C. (2004). *Character strengths and virtues: A handbook and classification*. Cary, NC: Oxford University Press, Inc.

Sa, W. C., Stanovich, K. E., & West, R. W. (1999). The domain specificity and generality of belief bias: searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*, 497-510.

Stanovich, K. E., & West, R. W. (1997). Reasoning independently from prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342-357.

Suedfeld, P., & Tetlock, P. (1977). Integrative Complexity of Communications in International Crises. *Journal of Conflict Resolution, 21*, 169-184.

Williams, K., Wise, S. L., & West, R. F. (2001, April). *Multifaceted measurement of critical thinking skills in college students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Willis, G. B. & Lessler, J. T. (1999). *Question Appraisal System BRFSS-QAS. A guide for Systematically Evaluating Survey Question Wording*. Prepared for BRFSS Annual meeting, May 1999.

Table 1

WCCES Items, Source, and Descriptive Statistics (N = 1001)

Item #	Item	Source ^a	Mean ^b	SD	Skewness	Kurtosis
1	It is important to persevere in your beliefs even when evidence is brought to bear against them.(R)	BI	2.63	1.12	-0.48	0.05
2	Certain beliefs are just too important to abandon no matter how good a case can be made against them.(R)	BI	3.12	1.36	-0.33	-0.56
3	One should disregard evidence that conflicts with your established beliefs.(R)	BI	4.28	1.14	0.49	-0.2
4	People should always take into consideration evidence that goes against their beliefs	FT	4.52	1.17	-0.82	0.48
5	Beliefs should always be revised in response to new information or evidence.	BI	3.63	1.19	-0.21	-0.24

Note. (R) indicates that item is reverse-scored.

^a BI = Belief Identification Subscale, Sá et al. (1999); FT = Flexible Thinking Subscale, Stanovich and West (1997)

^b Means computed after reverse scoring of items

Table 2

Item Correlations After Reverse-Scoring

Item #	1	2	3	4
1	1			
2	0.32	1		
3	0.29	0.25	1	
4	0.13	0.12	0.28	1
5	0.13	0.15	0.10	0.25

Table 3

Coding Framework

Organizing Theme	Basic Theme	Codes	Example	Item
Lack of Clarity	Grammar/Style	Confused	This is confusing. I am not sure what that means. Re-reads the question. They are tricky. Sometimes I wonder if I am giving conflicting views.	1, 2, 3, 5
	Multiple Interpretations	Multiple definitions, meanings, interpretations	Keeping your belief - that's one thing but considering it is a different. When you mean – consideration – do you mean – thinking about the evidence or incorporating it into your belief? They could go a lot of different ways, like your definition of “beliefs your definition of “consideration of evidence”	1, 2, 3, 4, 5
	Condition-Dependent	Circumstantial It depends	Depends on a belief and depends on the evidence. Loyalty depends on the circumstances. A lot of things are based on circumstances. My answer to every one of them depended on what specifically I was thinking about.	1, 2, 3, 4, 5
Socially Desirable Response Set	Positive connotations	Positive	Persevering in your beliefs is noble. Beliefs are a core of someone is. It is a solid foundation of your personality.	1
	Negative connotations	Negative	Whenever I read the word “abandon”, I think about abandoning children. Disregarding evidence completely is just ignorance.	2, 5
Construct Irrelevancy	Religion	Religion	The word belief makes them think about religion. Stories and examples of religious beliefs and contradictory evidence.	1, 2, 3, 4, 5
Inadequate Range of responses		Scale	Undecided would be a good one. I wish there was a middle one “I don’t know”. But if it was there, everyone would choose that. Neither agree nor disagree.	1, 2, 3, 4, 5

Figure 1. Scree plot and parallel analysis results.

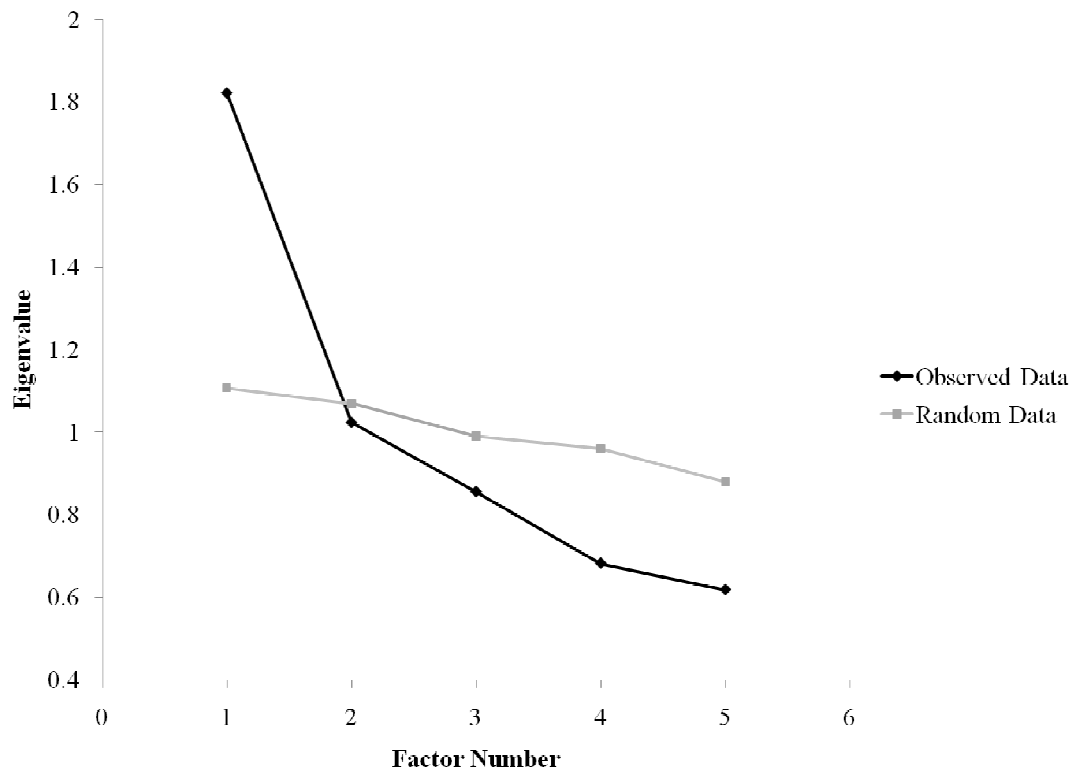


Figure 2. Thematic network of students' responses to the WCCES.

