The Acceptance and Action Questionnaire – II: An Item Response Theory Analysis

Clarissa W. Ong*[a]

Benjamin G. Pierce*[a]

Douglas W. Woods [b]

Michael P. Twohig [a]

Michael E. Levin [a]

[a] Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, UT 84322-2810

[b] Department of Psychology, Marquette University, 1250 W. Wisconsin Ave., Milwaukee, WI 53233

*Co-first authors

Funding: This research was supported by the NIMH of the National Institutes of Health under Award number R01MH080966 (Woods; PI).

Conflict of Interest: The authors declare that they have no conflict of interest.

Abstract

Psychological flexibility is the act of being open to internal experiences while pursuing valued life directions and has been implicated in positive mental health. A lack of psychological flexibility has been implicated in a wide range of mental health problems. In most research, assessment of psychological (in)flexibility has been done with the Acceptance and Action Questionnaire – II (AAQ-II), yet researchers have noted that items on the AAQ-II may not adequately discriminate between responses to experiences and the experiences themselves. Furthermore, little research has examined whether items on the AAQ-II function as intended in terms of assessing psychological (in)flexibility, and whether items function differently across populations. The present study used an item response theory framework to examine item functioning in the AAQ-II across items (within the measure) and across non-distressed student, distressed student, outpatient, and residential samples. The analyses identified differences in functioning between items, with some items being more sensitive to differences in psychological inflexibility. No items performed well in assessing psychological flexibility (as opposed to inflexibility) or positive functioning. Items functioned similarly across samples, yet patterns of responding differed in the non-distressed student versus residential and outpatient samples. Implications for the use of the AAQ-II in clinical and research contexts are discussed.

Keywords: psychological flexibility, assessment, psychometric, item response theory

The Acceptance and Action Questionnaire – II: An Item Response Theory Analysis

Acceptance and commitment therapy (ACT) is an acceptance- and mindfulness-based intervention situated in the third wave of cognitive-behavioral therapies (Hayes, 2004). ACT has been found to be effective in the treatment of physical and psychological presentations with comparable performance to established interventions (A-Tjak et al., 2015). The overarching aim of ACT is to improve psychological flexibility, which is the act of being open in an intentional manner to direct experiences as they occur and to engage in behaviors consistent with valued life directions (Hayes, Luoma, Bond, Masuda, & Lillis, 2006). Psychological flexibility is theorized to be a transdiagnostic construct implicated in general psychological wellbeing (Kashdan & Rottenberg, 2010), and that contributes to positive psychological functioning (Hayes et al., 2006). Psychological flexibility can be broken down into six core processes: acceptance, defusion, contact with the present moment, self as context, values, and committed action.

Acceptance refers to adopting an active, welcoming stance toward internal experiences without attempting to alter their frequency or form. This means choosing to be open to uncomfortable sensations, thoughts, and feelings without trying to fight or control them in any way. Defusion refers to responding to a thought as a thought—an automatic verbal product of our mind—rather than as reality, or to a sensation as a sensation rather than a "negative" feeling. It entails viewing inner experiences for what they are, not what they say they are. Contact with the present moment means noticing events as they occur, without evaluating them. It entails flexibly attending to the here and now without ruminating on the past or worrying about the future. Self as context is a process of perspective taking, wherein the individual sees the self as a vantage point from which thoughts and feelings are observed, rather than the thoughts and feelings themselves. Simply put, one takes the perspective of a boundless context in which inner

experiences occur, much like a sky that views various weather elements, regardless of what they are. From this stance, the individual is not controlled by thoughts and feelings, which are viewed as a transient part of the self. Values are individually chosen, meaningful life directions or qualities of behavior that can be enacted in any given moment. Examples of values include integrity, activism, openness, and kindness. It is imperative that values have intrinsic meaning to the individual such that they acquire reinforcing functions. Committed action are behaviors linked to values. Together, these six processes comprise a skill set termed psychological flexibility.

Low levels of psychological flexibility reflect psychological inflexibility, which is characterized by experiential avoidance, cognitive fusion, preoccupation with the past and/or future, attachment to self-identity, lack of values clarity, and inaction/impulsivity (Hayes et al., 2006). Psychological flexibility and inflexibility represent anchors on a continuum and individuals can vary in their level of flexibility along this scale. Inflexible processes manifest as an unwillingness to be open to aversive internal experiences that are perceived as having the power to dictate behaviors and ruminating on the past or worrying about the future in ways that detract from leading a fulfilling life. Broadly, psychological inflexibility describes a pattern of rigid behavioral responses to internal experiences that interfere with the pursuit of valued domains — the opposite behavioral pattern to flexibility.

Acceptance and Action Questionnaire – II

The measure most commonly used to assess psychological inflexibility — particularly in the context of ACT research — is the Acceptance and Action Questionnaire – II (AAQ-II; Bond et al., 2011). The AAQ-II has been used across nonclinical and clinical samples in various study designs, including cross-sectional surveys in college students (e.g., Levin, MacLane, et al.,

2014), randomized controlled trials for specific psychological conditions (e.g., Juarascio, Schumacher, Shaw, Forman, & Herbert, 2015; Lappalainen et al., 2014), and laboratory experiments (e.g., Ritzert, Forsyth, Berghoff, Barnes-Holmes, & Nicholson, 2015). In addition, numerous condition-specific versions based on the AAQ have been validated, including for body image (Sandoz, Wilson, Merwin, & Kellum, 2013), cardiovascular disease (Spatola et al., 2014), chronic pain (McCracken, Vowles, & Eccleston, 2004), stigma (Levin, Luoma, Lillis, Hayes, & Vilardaga, 2014), and trichotillomania (Houghton et al., 2014). The prevalence with which the AAQ and its variations have been and are being used underscore the centrality of psychological flexibility as a construct of interest within the ACT literature; this is unsurprising given that increasing psychological flexibility is arguably the end goal of ACT. At the same time, the importance ACT researchers place on psychological flexibility as the primary — if not, ultimate — arbiter of clinical progress and theoretical coherence warrants a need for accurate, reliable measurement of the construct.

The AAQ-II was developed with the goal of creating a more psychometrically sound version of its predecessor, the nine- or 16-item AAQ (Hayes et al., 2004). Across six samples of 2,816 participants, consisting of students, employees, and individuals seeking psychological treatment for substance use from North America and Europe, the AAQ-II demonstrated factor structure stability, internal consistency, test-retest reliability, convergent validity, predictive validity, and discriminant validity (Bond et al., 2011). The AAQ-II appears to measure a unidimensional factor across varied samples, consistent with theory that suggests psychological inflexibility functions as a coherent construct. In addition, as predicted, the AAQ-II was associated with higher levels of depression, anxiety, stress, and overall psychological distress.

Despite its widespread use, researchers have noted limitations of the AAQ-II. Issues include questionable construct validity and weak discriminant validity (e.g., Gámez, Chmielewski, Kotov, Ruggero, & Watson, 2011; Wolgast, 2014). For example, one criticism is that the AAQ-II confounds psychological outcomes with the process of psychological inflexibility, which can lead to artificially high correlations between the AAQ-II and measures of psychological distress. Yet, the AAQ-II purports to measure *responses* to internal experiences and consistency of behavior with values, not levels of aversive psychological stimuli. As a result, the AAQ-II shows poor discriminant validity, overlapping with items intended to measure distress (e.g., "I often feel depressed, worried or anxious") more so than with items designed to measure psychological inflexibility processes (e.g., "I let my thoughts and feelings come and go, without trying to control or avoid them;" Francis, Dawson, & Golijani-Moghaddam, 2016; Wolgast, 2014). Weak discriminant validity may be due to inaccurate operationalization of psychological inflexibility in the scale or issues with specific items (e.g., unclear wording).

Furthermore, although preliminary evidence supports the use of the AAQ-II for divergent samples (Bond et al., 2011), the extent to which the AAQ-II functions (i.e., is interpreted and answered) similarly across the wide range of populations in which it has been used remains unclear and has not been empirically verified. For instance, the only clinical sample included in the initial validation article was 290 individuals seeking treatment for substance use, yet the AAQ-II has been used in samples as diverse as women with trauma-related concerns (Fiorillo, McLean, Pistorello, Hayes, & Follette, 2017) and women staying in a residential eating disorder treatment facility (Juarascio et al., 2015). These differences could substantially impact how individuals interpret and respond to items on the scale, leading to issues with internal consistency and limited generalizability of obtained scores.

Item Response Theory

Item response theory (IRT), or modern test theory, provides a means to examine the association between individuals' item-level responding and the underlying latent trait of interest, referred to as theta (θ). IRT measures θ and evaluates the properties of test items by fitting statistical models to item-level responding. It differs from classical test theory (CTT) in a few ways. First, IRT focuses on item-level responding to a greater extent than CTT, which tends to use test-level indices (Harvey & Hammer, 1999). As such, more information on items can be gleaned from an IRT analysis. Second, IRT does not assume that measurement precision is constant across the possible range of test scores, whereas CTT does (Harvey & Hammer, 1999). Because of this, measurement precision of a test is represented by a continuous function, rather than a static figure. That is, an IRT approach accounts for variability in measurement precision depending on the individual's test score. Third, IRT methods enable evaluation of bias in test items using differential item functioning (i.e., whether items function differently for different subgroups) as well as estimation of the cumulative impact of item biases on test score (Harvey & Hammer, 1999). This is particularly important for understanding how test bias influences the performance of various subgroups. On the other hand, CTT techniques are unable to attribute observed mean subgroup differences to test bias or to a true difference in the level of the underlying trait. Still, IRT is limited because it cannot provide an ontological assessment of what the AAQ-II actually measures. In this study, we assume that θ refers to psychological flexibility based on theory not empirical verification.

To date, psychometric properties intrinsic to the AAQ, such as reliability and factor structure, have been examined across samples, and they appear to be relatively consistent (Bond et al., 2011). IRT analyses additionally evaluate differential functioning of items within the scale,

as well as of items across responder characteristics, providing information on the utility of individual items on the AAQ and whether responses to items across populations are equivalent with respect to construct measurement. Moreover, because IRT analyses are conducted at the item-level, they allow for variation in item functioning and evaluation of relative effectiveness across items, rather than only examining the scale as a whole. IRT enables such investigations by creating a latent construct that explains the primary source of variance in scores, against which item functioning is compared.

Receiver Operating Characteristic Analyses

Receiver operating characteristic (ROC) analyses can supplement IRT-informed investigations. ROC curves provide an assessment of the validity of individual items or scale scores to discriminate among individuals meeting or not meeting an external cutoff point. Thus, if items are found to perform differently within the AAQ-II, their utility to discriminate among levels of other constructs can be evaluated via ROC curve estimation. Parameters from the IRT and ROC analyses can then be combined to identify items that most effectively detect differences in psychological inflexibility (per IRT parameters) and discriminate among an external criterion (based on ROC curve estimates).

Present Study

The purpose of the current study was to use an IRT approach to empirically evaluate the utility of the AAQ-II as a measure of psychological (in)flexibility vis-à-vis various populations without any a priori assumptions about its functional properties. Given our use of clinical samples, we have referred to the latent construct of interest as “psychological inflexibility” to communicate our focus on individuals with lower levels of psychological flexibility. However, as mentioned above, psychological flexibility and inflexibility are believed to reflect differing

levels of the same construct. In this study, we operationalized utility as discrimination strength (ability to detect differences in levels of the latent construct). In addition, we were interested in consistency of difficulty (threshold that reflects a certain level of the latent construct) across samples. Items with high discrimination strength are deemed to be more sensitive to varying levels of psychological inflexibility, whereas consistency of both discrimination ability and difficulty across samples indicate that items can be used with different populations and administrators can reasonably expect that equivalent scores reflect similar levels of psychological inflexibility.

We had four specific research questions. First, are items on the AAQ-II equally sensitive to varying levels of psychological inflexibility? Given that certain items of the AAQ are sensitive to constructs besides psychological inflexibility (Wolgast, 2014), and may be insensitive to changes in inflexibility, individuals may not be interpreting or responding to these items consistently with how they were designed. In other words, while the scale may be assessing psychological inflexibility, some items may not be as sensitive to detecting variations in this construct as we expect. An IRT analysis evaluates the performance of items in relation to the latent construct (i.e., psychological inflexibility) that is presumably responsible for covariance across the items and is able to detect which items are more or less sensitive to that common source of variability. Second, do the same item scores reflect similar levels of psychological inflexibility? This question concerns how "difficult" it is (or the level of psychological inflexibility needed) to obtain a specific score on an item, and addresses issues related to over- or underreporting of psychological inflexibility across items. For example, a response of 3 on an "easier" item might reflect a lower level of psychological inflexibility than a response of 3 on a more "difficult" item. By extension, a total score of 40 might represent a similar level of

psychological inflexibility to a total score of 30. Because IRT focuses on actual patterns of responding relative to the latent construct as opposed to expected patterns of responding, such analyses might clarify how well obtained scores match up with level of psychological inflexibility, improving our interpretation of AAQ-II scores. Third, do items function differently depending on responder characteristics? Research to date assumes that the AAQ-II can be administered to various populations, and scores obtained are comparable across groups (Bond et al., 2011). However, as outlined in the previous section, CTT does not provide a robust test of inter-population reliability. IRT analyses can determine if differences in item sensitivity and "difficulty" across groups are attributable to sample characteristics (e.g., presence of psychopathology, interpretation of items; test bias), or to degree of psychological inflexibility. Distinguishing between these two sources of variance has implications for test interpretation and use of clinical cutoffs across samples, as inconsistent responding to items across samples may influence the validity and interpretability of AAQ-II scores in these groups. Fourth, what do total and item scores on the AAQ-II tell us about "clinically significant" psychological inflexibility? Bond et al. (2011) recommended a cutoff between 24 and 28 on the AAQ-II total score to indicate presence of clinically significant distress based on results from a regression model with AAQ-II total score as the dependent variable and established cutoffs on measures of psychological distress (e.g., Beck Depression Inventory-II [BDI-II]) as predictors. This method of evaluating a clinical cutoff has two limitations: (1) focusing only on the total score obscures the effects of differential item functioning (i.e., a score of 26 may not always reflect the same level of psychological inflexibility); and (2) ROC curves have been used more commonly to determine diagnostic thresholds, and may be a more appropriate statistical technique to answer this question (Fan, Upadhye, & Worster, 2015; Greiner, Pfeiffer, & Smith, 2000).

Method

Participants

The present sample comprised three subsamples. The first subsample included 1,146 students from a Western college in the United States who completed online surveys between the years 2014 and 2017. The second subsample included 111 women from a residential treatment center for eating disorders, who met DSM-V criteria for at least one eating disorder. The third sub-sample included 90 adults meeting criteria for trichotillomania who were recruited for a randomized controlled trial of an outpatient treatment. All samples completed the AAQ-II and provided demographic information. The student sample completed an additional measure of symptoms and functional impairment. Informed consent was obtained from all individuals included in the study.

Measures

Demographic items. All participants provided their age, gender, race, and ethnicity. Age was entered numerically, whereas gender, race, and ethnicity were selected from multiple-choice options. The demographic items were used to characterize the samples.

The Acceptance and Action Questionnaire – II (Bond et al., 2011). The present study examined the item-response characteristics of the 7-item version of the AAQ-II (a 10-item version is available; see Bond et al., 2011). Items are rated on a 7-point scale ranging from 1 = never true to 7 = always true. Preliminary evidence suggests that the AAQ-II has adequate reliability and validity in non-clinical (e.g., college students) and clinical samples (Bond et al., 2011; Fledderus, Oude Voshaar, Ten Klooster, & Bohlmeijer, 2012). The AAQ-II, and the original AAQ, are by far the most commonly used measures of psychological inflexibility in research to date (Hayes, Levin, Plumb-Vilardaga, Villatte, & Pistorello, 2013).

Psychological symptoms. The Counseling Center Assessment of Psychological Symptoms, 34-item version (CCAPS-34; Locke et al., 2012) was used to assess a range of mental health concerns in the college student sample. The CCAPS-34 includes subscales for depression, generalized anxiety, social anxiety, academic distress, eating concerns, hostility, and alcohol use. All items are rated on a 5-point scale ranging from 0 = not at all like me to 4 = extremely like me, and clinical significance is determined using the sum of items across each subscale. The CCAPS-34 has been found to have adequate reliability and validity in previous studies with college students (Center for Collegiate Mental Health [CCMH], 2012).

Analytical Plan

IRT analyses were used to examine item functioning across the student, residential, and outpatient samples. To reduce unexplained heterogeneity in these models, the student sample was split into two subgroups, consisting of students who exceeded at least one clinical cutoff on the CCAPS-34 (elevated subsample) and students without any clinical elevations on the CCAPS-34 (normative subsample). Multi-group models were used to assess item functioning within the AAQ-II across the normative and elevated student subsamples, the residential subsample, and outpatient subsample. Parameter constraints were used to determine whether items functioned similarly across these subgroups.

The graded response model (GRM; Samejima, 1997) was used to assess scale and item-level functioning. The GRM is an extension of the two-parameter logistic IRT model for scales with polychotomous item choices. For each item, the GRM examines a series of dichotomies between responses less than a given point on a scale and responses equal to or greater than that point. Based on these dichotomies, the GRM computes a series of "difficulty" and discrimination parameter for that item. The "difficulty" parameters assessed at what level of the measured

construct 50% of the sample would be expected to score equal to or above a given point on a scale, as compared with the value adjacent to it. The discrimination parameter for each item provided an overall assessment of how well an item measured variability in the latent construct, independent of the levels at which this construct were assessed. Based on these parameters, a total information function (i.e., a curve) for each item was computed to describe the accuracy of the item's performance relative to levels of the latent construct.

Each research question can be examined based on specific parameters of the GRM. The discrimination parameters provide an assessment of the sensitivity of each item relative to differences in level of psychological inflexibility, with larger discrimination values reflecting greater sensitivity. The "difficulty" parameters provide an assessment of the degree of psychological inflexibility reflected by a given response to a given item and were used to evaluate differences in the levels of inflexibility assessed by responses across items (e.g., whether a response of "3" reflects similar levels of inflexibility across two items). Finally, the information functions were used to assess the relative contributions of each item to variance in psychological inflexibility, thus providing an assessment of the "relevance" of each item within the scale.

The GRM assumes that items function similarly across groups, therefore, differential item functioning was evaluated across the sub-samples prior to interpreting the discrimination and "difficulty" parameters as well as relative information contributed by each item. Differences in these parameters among the sub-samples were assessed by constraining the discrimination or "difficulty" parameters to be equal across groups. The adjusted Bayesian Information Criterion (aBIC) was used to evaluate differences in model fit, as this index has demonstrated precision in differentiating model fit among competing GRM models (Kang, Cohen, & Sung, 2009). In

addition, inspection of the residuals associated with item response categories was used to assess local areas of misfit in groups that are not detected by global differences in the aBIC.

Receiver Operating Characteristic Curve Analyses

In addition to the GRM analyses, receiver operating characteristic (ROC) curve analyses were used to provide further interpretation of item and scale-level functioning of the AAQ-II. ROC curves plot the proportion of the sample classified as true-positive and false-positive on a known dichotomous outcome (e.g., "not clinically significant" vs. "clinically significant"), based on increasing values of a measure. Based on the ROC plot, optimal "cutoff" scores that most accurately classify participants can be identified. Sensitivity is defined as the proportion of participants correctly classified as positive on the known outcome. Conversely, specificity is defined as one minus the proportion of participants incorrectly classified as positive on this outcome.

Information on sensitivity and specificity can be combined with the results of item response theory analyses to suggest possible interpretations of item and scale-level AAQ-II scores. The ROC curve analyses were performed at both the item and scale-level, assessing the sensitivity and specificity of item scores and the total scale score to discriminate between participants in student sample with and without elevated CCAPS-34 scores. The outpatient trichotillomania and residential eating disorder samples were not included in the ROC analyses because of the specificity of these presenting concerns.

Results

Descriptive statistics were used to characterize the study samples prior to analyses. The mean age in the college sample was 21.1 years (SD = 5.6, range = 18.0 to 55.0), with gender distributed as 63.9% women, 36.0% men, and 0.1% other gender. The college sample was

mostly White (94.2%), with 2.4% Asian, 1.4% Black, 4.3% Latinx/Hispanic, 1.5% Native American, 0.7% Native Hawaiian or Pacific Islander, and 1.6% "other" ethnicity. Twenty-two percent of the college sample had at least one clinically elevated subscale on the CCAPS; of those with elevated scores, the most common problems among students were eating-related concerns (23.6%) depression (23.1%), and anxiety (22.8%). The mean age of the residential sample was 23.81 years (SD = 6.61, range = 18.0 to 54.0). Ninety-two percent of the residential inpatient sample identified as White, 1.2% as Asian, 1.2% as Black/African-American, 2.4% as Latinx/Hispanic, and 3.6% as "other" ethnicity. The modal education in this sample was a high school degree (51.4%), with 29.7% completing a college degree (Bachelor's or Associate's), 9.0% not completing high school, and 4.5% completing a post-graduate degree (refer to Lee, Smith, Twohig, Lensegrav-Benson, & Quakenbush-Roberts, 2017 for more detail). The trichotillomania outpatient sample had an average age of 34.8 years (SD = 12.8, range = 18 to 61 years), with 92.2% women and 7.8% men, and 83.3% who identified as White (Houghton et al., 2014).

Graded Response Models

Differential item functioning. Three GRMs (Samejima, 1997) were fitted to the normative student, elevated student, residential, and outpatient samples to assess differential item functioning. The first model constrained the variance of psychological inflexibility to one across the four groups and included freely estimated discrimination and "difficulty" parameters across groups. The fit of this model was compared to a more restrictive model with the discrimination parameters set equal across groups. The model with equality constraints on the discrimination parameters showed improved relative fit to the data ($aBIC_1$ = 30779.91 versus $aBIC_2$ =

30733.34). This result indicated that invariant discrimination parameters across groups increased model parsimony.

The third GRM constrained both discrimination and "difficulty" parameters, specifying them as equal across groups. This model was compared to the second GRM that constrained only discrimination parameters, and showed improved relative fit ($aBIC_2$ = 30733.14 versus $aBIC_3$ = 30512.30), showing greater parsimony when both discrimination and "difficulty" parameters were invariant across groups. Hence, items on the AAQ-II appeared to perform similarly across the normative student, elevated student, residential, and outpatient subgroups.

Because the model with restrictions on both discrimination and "difficulty" parameters offered the best relative fit to the data, the results of this model were further interpreted to assess the equivalence of items as indicators of psychological inflexibility. The results of this model are presented in Table 1, which presents the discrimination and "difficulty" parameters.

Item discrimination. Item discrimination parameters were interpreted to assess differences in item sensitivity to variations in psychological inflexibility. In Table 1, items 3 and 4 showed the largest discrimination parameters, indicating these items were most sensitive to individual differences in psychological inflexibility. Conversely, items 2 and 6 appeared less sensitive to individual differences in the latent psychological inflexibility variable and to share the least in common with other items in the scale.

Findings on the discrimination parameters suggest that the AAQ-II items differ in sensitivity to individual differences. These findings also suggest that changes in certain items may be more meaningful (i.e., clinically relevant) than others. An important pattern in these findings is that items that specify the function of an internal experience, for example, "my

painful memories prevent me from living a fulfilling life," seem to be more sensitive than generally worded items such as "it seems like most people are living their lives better than I am."

Item difficulty. Item "difficulty" parameters were used to assess differences in the levels of psychological inflexibility reflected by various responses to the items. Inspection of the "difficulty" parameters suggests that responses above 5 on items 1, 3, and 4 reflect especially high levels of psychological inflexibility. Conversely, elevated responses on items 6 and 7 may be obtained with lower psychological inflexibility scores. These findings suggest that higher scores on items 1, 3, and 4 may be more clinically relevant than higher scores on items 6 and 7. For instance, a score of 4 on item 6 corresponds to a level of psychological inflexibility that may be over 2 SDs lower that the level of inflexibility suggested by a score of 4 on item 4.

Differences in the item "difficulty" parameters suggest that the meaning of a respondent's total AAQ-II score depends on which items were especially elevated. Certain items (e.g., items 1, 3, and 4) may describe behavior associated with very high levels of psychological inflexibility, whereas other items (e.g., items 6 and 7) may describe behavior associated with more moderately inflexible responses.

Item and scale information functions. The total information function for all seven AAQ-II items across sub-groups is presented in Figure 1. As illustrated, the AAQ-II items provided the most information at higher levels of psychological inflexibility (i.e., scores at least 0.5 SD above the sample mean). Inspection of individual item information curves revealed that items 3 and 4 contributed the most information, yet mostly at higher levels of inflexibility; conversely, none of the items performed well at lower levels of inflexibility (i.e., at 1 SD below the latent mean in the normative student sample).

Findings on the information contributed by each item and the scale thus suggest that the AAQ-II items more accurately measure psychological *inflexibility* than psychological *flexibility* (i.e., low levels of psychological inflexibility) across the different samples. Further, these findings indicate that responses to items 3 and 4 provide the best indicators of behavior assessed among all the AAQ-II items.

Assessment of Model Residuals

Qualitative analyses were performed on the residuals of the GRM with invariant discrimination and "difficulty" parameters across groups (i.e., the best-fitting model) to explore group differences besides patterns of differential item functioning. While the aBIC is sensitive to global misfit across groups, it does not identify areas of local misfit wherein an item or specific response category may function poorly for a specific group of participants. Thus, inspection of the residuals plots was used to determine if certain responses to the items were systematically over- or under-represented in certain groups when item functioning was assumed to be equivalent (per the discrimination and "difficulty" parameters in the GRM). The residuals were measured as the difference between observed and expected proportions of responses in each category for each item. A plot of the mean absolute residual across response categories for each AAQ-II item for all four groups is presented in Figure 2; this plot combines over and under-estimation of response frequencies across the seven response categories, using absolute discrepancies, providing an estimate of gross error associated with each item. A plot of the mean residual across AAQ-II items for each response category for all groups is presented in Figure 3; this plot illustrates the tendencies of each response category (the categories from 1 = *never true* to 7 = *always true*) to be over- or under-represented in each participant group.

As illustrated in both figures, the GRM showed little evidence of local misfit within the elevated student group, whereas residuals were higher across items and response categories for the normative student, residential, and outpatient groups. As Figure 2 shows, the residential and outpatient groups had higher residuals across most AAQ-II items, with the greatest absolute residuals observed for item 1 in the residential group. Figure 3 illustrates that the residential and outpatient sample tended to respond at extreme values (1 or 7) more frequently than expected based on the GRM, and provided mid-range responses less frequently than expected. Conversely, the normative student sample provided fewer responses at the lowest value 1 = never true (in the direction of lower psychological inflexibility) than expected based on the model and provided more mid-range responses.

These findings suggest that inpatients may adhere to a more extreme pattern of responding, which may result in inflated or deflated scores, whereas non-distressed students may be unwilling to endorse a lower extreme, which may inflate scores among those with very low psychological inflexibility. While the results are primarily descriptive, the present analyses suggest that certain AAQ-II items and response categories may elicit slightly different response patterns depending on problem severity and population assessed.

ROC Curve Analysis

The ROC analyses suggested total AAQ-II scores between 28 and 32 maximized sensitivity and specificity in discriminating between students with and without elevated CCAPS-34 scores. A score of 28 maximized sensitivity, with adequate levels of specificity (sensitivity = 0.903, specificity = 0.807) while a score of 32 maximized specificity while retaining sufficient sensitivity (sensitivity = 0.766, specificity = 0.897). This range of scores was selected based on the combinations of sensitivity + specificity values that provided the highest sums (range = 1-2),

without allowing either to fall below a threshold of 0.750 (i.e., 25% false-positives or false-negatives). At the item level, a cutoff of 4 = *sometimes true* on items 1, 3, and 4 maximized both sensitivity and specificity. A cutoff of 5 = *frequently true* maximized these values on items 2, 5, 6, and 7. However, as indicated by the GRM analyses, items 3 and 4 provided the most information. Therefore, these findings suggest that a lower-bound cutoff of 28 may most effectively classify students with elevated CCAPS-34 scores in the present sample, if this score also includes a response to item 3 or 4 at "4" or above. Conversely, a higher-bound cutoff of 32 seems to provide more effective classification in the absence of elevations on these items.

Discussion

The present study evaluated the psychometric properties of the AAQ-II using IRT as a framework for investigating item functioning as well as differences in item functioning across student, residential, and outpatient samples. Results across samples suggested that certain items were more effective than others in discriminating among levels of psychological inflexibility. Namely, items 3 and 4, which asked more specifically about the functions of internal events (e.g., "I worry about not being able to control my worries and feelings") appeared to show greater discrimination at moderate and elevated levels of psychological inflexibility. Conversely, more generally worded items, such as, "It seems like most people are living their lives better than I am," showed weaker discrimination overall with a slightly higher ability to differentiate very low levels of psychological inflexibility.

A failure to attend to differences in the information contributed by items may lead to incorrect conclusions if the AAQ-II items are simply summed and interpreted in a total score. Attending to client responses to items 3 and 4 may provide the most clinical utility, because these items seem to be most sensitive to differences in psychological inflexibility. Further,

elevated scores on these items appear to reflect an especially high degree of inflexibility, such that may assist in detecting clients who require a more intensive intervention. Conversely, other items may provide less information because they are broadly worded (e.g., “It seems like most people are living their lives better than I am”) and may reflect patterns of responding that are more characteristic of the general population. Individuals whose primary elevations are on items 2, 6, or 7 may thus benefit from a less intensive intervention as compared with those who score highly on items 3 and 4.

Qualitative data can be useful to clarify the meanings of items to individual participants. For example, for item 4, clinicians may ask how painful memories get in the way of living a meaningful life or what answering “frequently true” on that item means to clients. Of note, our results do not necessarily imply that the item scores per se are indicative of the level of latent trait of psychological inflexibility; it is the behavioral response to these items that reflect the latent construct. For both research and clinical reasons, it could also be helpful to explore how individuals interpret these items, which may facilitate identification of themes that more powerfully measure psychological inflexibility.

In general, all items provided little information at very low levels of psychological inflexibility (i.e., more than 0.5 SD below the mean; or higher levels of psychological flexibility). The steep drop-off in information provided by items as the total score decreases suggests that AAQ-II total scores may not be interpretable unless respondents score at least 17 (based on the college student sample) — given that inflexibility scores below this cutoff were not well-detected or discriminated by any item. Therefore, even though low-scoring data are usable, they may not be useful. The loss of information was more pronounced in the college student sample, and so extra caution should be exercised when working with low-scoring AAQ-II data in

groups with similar characteristics. Low scores should not be interpreted as accurately reflecting psychological inflexibility, and an alternative measure seems necessary to assess more flexible repertoires of behavior.

Conversely, items of the AAQ-II provided the most information at higher levels of inflexibility. The ROC analyses suggested a "clinical cutoff" between 28 and 32 for best discriminating between students with and without elevated symptoms. Scores at the lower end of this range may provide the most effective classification if they are accompanied by elevations on items 3 or 4. Altogether, combining item and scale-level information in interpreting the meaning of AAQ-II scores may serve to mitigate problems associated with differences in discrimination, "difficulty," and the amount information provided by different items. Of note, the "cutoff scores" and their interpretation presented here should be used tentatively due to the homogeneity of the student sample as well as the way that the participants were classified (i.e., solely based on CCAPS-34 scores). Furthermore, this cutoff is more conservative than that recommended by Bond et al. (2011), who used different outcomes measures (e.g., Beck Depression Inventory-II, General Health Questionnaire-12) to characterize psychological distress. More data are needed to determine a clinical cutoff on the AAQ-II that can be broadly applied, if one exists, across various outcomes and populations.

The multigroup analyses revealed similar psychometric properties across students with and without elevated CCAPS-34 scores, outpatients with trichotillomania, and residential patients with eating disorders. A model with identical discrimination and "difficulty" parameters fit the data well across groups. Thus, the AAQ-II items did not show differential functioning based on symptom severity or clinical presentation. In other words, changes in scale scoring (e.g., three-point increase) are likely to be equivalent across samples and effect size estimates

can be reliably compared across samples. Thus, the AAQ-II is likely suitable for use in clinical and non-clinical samples.

However, an inspection of residuals across these groups suggested some differences in the accuracy of information provided by individual items (Figure 2). Specifically, items 2 ("I'm afraid of my feelings") and 7 ("worries get in the way of my success) performed worse in the outpatient sample than other samples, whereas item 1 ("my painful experiences and memories make it difficult to live a life that I would value") performed poorly in the residential sample. These patterns may be due to the contexts (e.g., residential versus outpatient) as well as the kinds of problems (e.g., trichotillomania versus eating disorders). Clinicians should consider the ways that clients respond to the AAQ-II items based on their contexts; for example, clients who do not present with significant anxiety may not endorse items 2 and 7 as highly because these items ask specifically about experiences of fear and worry. On the other hand, individuals in a residential setting may have problems envisioning "a life they would value" per item 1, therefore this item may not effectively assess inflexible responding. Our results lead to a similar conclusion to that from other studies that have included both the AAQ and domain-specific AAQ in their explanatory models. These studies tend to find that domain-specific measures do not overlap perfectly with the AAQ-II and are more strongly linked to outcomes relevant to their samples (e.g., Houghton et al., 2014; Sandoz et al., 2013; Spatola et al., 2014), which suggest that domain-specific AAQs could measure inflexibility in more content-valid ways than the AAQ-II, and provide a unique contribution in terms of assessment of psychological inflexibility. Thus, it may be helpful to use both the AAQ-II and domain-specific AAQs to obtain a more comprehensive picture of psychological inflexibility.

Further, the discrepant mean residual pattern averaged across items, presented in Figure 3, indicates that interpretation of extreme responses in clinical samples may need to be moderated. More extreme patterns of responding were observed in the residential and outpatient samples (in either direction), and asymptomatic students were less willing to endorse especially low levels of inflexibility. In other words, scores of 1 may not accurately reflect extremely low levels of psychological inflexibility in certain samples, and scores of 7 may not accurately reflect extremely high levels of psychological inflexibility in others. This means that clinicians should be conservative when interpreting extreme responses in either direction when working with clinical samples. In addition, due to the underreporting of very low levels of inflexibility in the college sample and poor overall sensitivity of items in this range, caution should be used when interpreting data from non-clinical and general college samples. Lastly, as psychological inflexibility is a relatively new target within clinical settings, it is important to be conscious of patterns of responding that may influence scores on items and measures used to assess this construct.

Despite demonstrating satisfactory psychometric properties based on classical test theory (CTT) methods, the AAQ-II does not appear to perform as well using IRT analyses. For instance, we found that the same total score may reflect different degrees of psychological inflexibility depending on the items endorsed and to some extent on the responder's characteristics. Hence, the AAQ-II may need to be modified in order to provide a more reliable assessment of psychological inflexibility – both across items within the scale as well as across samples to whom the scale is administered. A preliminary examination of item performance revealed that items containing broad language, such as "feelings" or "emotions," did not perform as well as items that used more specific terms such as, "worries" and "memories." Furthermore, items that

addressed the function of emotions (e.g., “I worry about not being able to control my worries and feelings”) or their impact (e.g., “My painful memories prevent me from having a fulfilling life”) provided the most information overall. Thus, using more precise language and explicitly asking about the function and/or effect of internal experiences may improve the utility of the AAQ-II.

In addition, our findings point to specific ways in which measurement with the AAQ-II may be streamlined and improved. First, it is possible that an abbreviated version of the AAQ-II may be useful for measuring psychological inflexibility. Our findings suggested that psychological inflexibility could be assessed in an internally consistent way with as little as three items, and that these items could contribute sufficient information to detect meaningful differences among individuals and possibly change over time. However, further investigation of the best-performing items (items 1, 3, and 4) and their functioning across a broader range of samples and over time is needed to support this assertion, before this suggestion is put into practice. Second, our findings suggest that revisions could be made to the AAQ-II to improve the consistency of items and their informational value across groups. Such improvements may help mitigate concerns about the (in)validity of items and their capacity to detect meaningful differences among individuals and across time.

Conversely, given the potential varying meaning of item and total scores on the AAQ-II, using other measures of psychological inflexibility, such as the Brief Experiential Avoidance Questionnaire (BEAQ; Gámez et al., 2014) and Comprehensive Assessment of Acceptance and and Commitment Therapy Processes scale (CompACT; Francis et al., 2016), alongside the AAQ-II might collectively provide a more accurate measure of psychological inflexibility. At the same time, because these measures have not been validated using IRT, further evaluation of these measures is needed as well.

Summarily, our findings show that IRT yields a different appraisal of measures from CTT. Thus, future psychometric research may benefit from using IRT analyses in addition to CTT methods in order to obtain a more holistic assessment of scale and item functioning. After all, it is difficult to make inferences about findings if we are uncertain about the reliability and validity of the very tools we use to measure amorphous constructs, such as psychological inflexibility. In this regard, quality assessment is essential for producing accurate findings that can then be translated to practice, theory, and continued research. Especially as researchers, if we wish to draw impactful conclusions from our data (e.g., whether or not a treatment is recommended for individuals struggling with a particular mental health concern), it is incumbent on us to ensure that our instruments are psychometrically sound and reflect the constructs we discuss. Using multiple methods to evaluate the psychometric properties of measures brings us closer to this goal.

Limitations

Our nonclinical samples solely comprised college students whose demographic homogeneity (e.g., majority White and younger than 25) limits generalizability of our results to other nonclinical populations. Similarly, our clinical samples consisted of individuals seeking treatment for specific concerns (i.e., trichotillomania, eating disorders). Although these samples allowed us to investigate item functioning of the AAQ-II across distinct contexts and client groups, they may limit the generalizability of our findings to other clinical populations. Replication of our findings in more diverse groups — particularly with regard to demographic variables and clinical presentation — is needed. Testing functioning of the AAQ-II across a wider range of subgroups would provide data on cultural or diagnostic specificity in terms of how individuals understand and respond to items on the AAQ-II. Such information is important

in determining how and when the AAQ-II should be administered and interpreted, depending on responder characteristics.

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

A-Tjak, J. G. L., Davis, M. L., Morina, N., Powers, M. B., Smits, J. A., & Emmelkamp, P. M. (2015). A meta-analysis of the efficacy of acceptance and commitment therapy for clinically relevant mental and physical health problems. *Psychother Psychosom, 84*(1), 30-36. doi:10.1159/000365764

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., . . . Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: a revised measure of psychological inflexibility and experiential avoidance. *Behav Ther, 42*(4), 676-688. doi:10.1016/j.beth.2011.03.007

Fan, J., Upadhye, S., & Worster, A. (2015). Understanding receiver operating characteristic (ROC) curves. *Cjem, 8*(01), 19-20. doi:10.1017/s1481803500013336

Fiorillo, D., McLean, C., Pistorello, J., Hayes, S. C., & Follette, V. M. (2017). Evaluation of a web-based acceptance and commitment therapy program for women with trauma-related problems: A pilot study. *Journal of Contextual Behavioral Science, 6*(1), 104-113. doi:10.1016/j.jcbs.2016.11.003

Fledderus, M., Oude Voshaar, M. A., Ten Klooster, P. M., & Bohlmeijer, E. T. (2012). Further evaluation of the psychometric properties of the Acceptance and Action Questionnaire-II. *Psychological Assessment, 24*(4), 925-936. doi:10.1037/a0028200

Francis, A. W., Dawson, D. L., & Golijani-Moghaddam, N. (2016). The development and validation of the Comprehensive assessment of Acceptance and Commitment Therapy processes (CompACT). *Journal of Contextual Behavioral Science, 5*(3), 134-145. doi:10.1016/j.jcbs.2016.05.003

Gámez, W., Chmielewski, M., Kotov, R., Ruggero, C., Suzuki, N., & Watson, D. (2014). The Brief Experiential Avoidance Questionnaire: Development and Initial Validation. *Psychological Assessment, 26*(1), 35-45. doi:10.1037/a0034473

Gámez, W., Chmielewski, M., Kotov, R., Ruggero, C., & Watson, D. (2011). Development of a Measure of Experiential Avoidance: The Multidimensional Experiential Avoidance Questionnaire. *Psychological Assessment, 23*(3), 692-713. doi:10.1037/a0023242

Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine, 45*, 23-41. doi:10.1016/S0167-5877(00)00115-X

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27*(3), 353-383. doi:10.1177/0011000099273004

Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy, 35*(4), 639-665. doi:10.1016/S0005-7894(04)80013-3

Hayes, S. C., Levin, M. E., Plumb-Vilardaga, J., Villatte, J. L., & Pistorello, J. (2013). Acceptance and commitment therapy and contextual behavioral science: examining the progress of a distinctive model of behavioral and cognitive therapy. *Behav Ther, 44*(2), 180-198. doi:10.1016/j.beth.2009.08.002

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: model, processes and outcomes. *Behav Res Ther, 44*(1), 1-25. doi:10.1016/j.brat.2005.06.006

Hayes, S. C., Strosahl, K., Wilson, K. G., Bissett, R. T., Pistorello, J., Toarmino, D., . . . McCurry, S. M. (2004). Measuring experiential avoidance: A preliminary test of a working model. *The Psychological Record, 54*, 553-578. doi:10.1007/BF03395492

Houghton, D. C., Compton, S. N., Twohig, M. P., Saunders, S. M., Franklin, M. E., Neal-Barnett, A. M., . . . Woods, D. W. (2014). Measuring the role of psychological inflexibility in Trichotillomania. *Psychiatry Res, 220*(1-2), 356-361. doi:10.1016/j.psychres.2014.08.003

Juarascio, A. S., Schumacher, L. M., Shaw, J., Forman, E. M., & Herbert, J. D. (2015). Acceptance-based treatment and quality of life among patients with an eating disorder. *Journal of Contextual Behavioral Science, 4*(1), 42-47. doi:10.1016/j.jcbs.2014.11.002

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*, 499-518. doi:10.1177/0146621608327800

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review, 30*(7), 865-878. doi:10.1016/j.cpr.2010.03.001

Lappalainen, P., Granlund, A., Siltanen, S., Ahonen, S., Vitikainen, M., Tolvanen, A., & Lappaleinen, R. (2014). ACT Internet-based vs face-to-face? A randomized controlled trial of two ways to deliver Acceptance and Commitment Therapy for depressive symptoms: An 18-month follow-up. *Behaviour Research and Therapy, 61*, 43-54. doi:10.1016/j.brat.2014.07.006

Lee, E. B., Smith, B. M., Twohig, M. P., Lensegrav-Benson, T., & Quakenbush-Roberts, B. (2017). Assessment of the body Image-Acceptance and Action Questionnaire in a female residential eating disorder treatment facility. *Journal of Contextual Behavioral Science, 6*(1), 21-28. doi:10.1016/j.jcbs.2016.11.004

Levin, M. E., Luoma, J. B., Lillis, J., Hayes, S. C., & Vilardaga, R. (2014). The Acceptance and Action Questionnaire - Stigma (AAQ-S): Developing a measure of psychological flexibility with stigmatizing thoughts. *J Contextual Behav Sci, 3*(1), 21-26. doi:10.1016/j.jcbs.2013.11.003

Levin, M. E., MacLane, C., Daflos, S., Seeley, J. R., Hayes, S. C., Biglan, A., & Pistorello, J. (2014). Examining psychological inflexibility as a transdiagnostic process across psychological disorders. *Journal of Contextual Behavioral Science, 3*, 155-163. doi:10.1016/j.jcbs.2014.06.003

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P. W., Hayes, J. A., Castonguay, L. G., . . . Lin, Y. C. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development, 45*, 151-169. doi:10.1177/0748175611432642

McCracken, L. M., Vowles, K. E., & Eccleston, C. (2004). Acceptance of chronic pain: component analysis and a revised assessment method. *Pain, 107*(1), 159-166. doi:10.1016/j.pain.2003.10.012

Ritzert, T. R., Forsyth, J. P., Berghoff, C. R., Barnes-Holmes, D., & Nicholson, E. (2015). The impact of a cognitive defusion intervention on behavioral and psychological flexibility: An experimental evaluation in a spider fearful non-clinical sample. *Journal of Contextual Behavioral Science, 4*(2), 112-120. doi:10.1016/j.jcbs.2015.04.001

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Sandoz, E. K., Wilson, K. G., Merwin, R. M., & Kellum, K. K. (2013). Assessment of body image flexibility: The Body Image-Acceptance and Action Questionnaire. *Journal of Contextual Behavioral Science, 2*(1-2), 39-48. doi:10.1016/j.jcbs.2013.03.002

Spatola, C. A., Cappella, E. A., Goodwin, C. L., Baruffi, M., Malfatto, G., Facchini, M., . . . Molinari, E. (2014). Development and initial validation of the Cardiovascular Disease Acceptance and Action Questionnaire (CVD-AAQ) in an Italian sample of cardiac patients. *Front Psychol, 5*, 1284. doi:10.3389/fpsyg.2014.01284

Wolgast, M. (2014). What Does the Acceptance and Action Questionnaire (AAQ-II) Really Measure? *Behavior Therapy, 45*, 831-839. doi:10.1016/j.beth.2014.07.002

Table 1

Results of the GRM with Equivalent Discrimination and "Difficulty" Parameters Across Groups

Item	*a*	b_2	b_3	b_4	b_5	b_6	b_7
1. My painful experiences and memories make it difficult to live a life that I would value	2.42	-0.09	2.23	3.32	5.02	6.81	8.35
2. I'm afraid of my feelings	2.17	-0.74	0.91	1.96	3.65	5.63	7.22
3. I worry about not being able to control my worries and feelings	2.62	-0.60	1.23	2.24	3.82	5.63	7.83
4. My painful memories prevent me from having a fulfilling life	2.71	1.08	2.95	4.07	5.64	7.24	9.13
5. Emotions cause problems in my life	2.23	-2.27	0.27	1.53	3.59	5.42	7.55
6. It seems like most people are living their lives better than I am	2.21	-2.60	-0.35	0.68	2.50	4.06	5.62
7. Worries get in the way of my success	2.29	-1.95	0.08	1.33	3.11	4.85	6.50

Note. a = discrimination parameter. b_x = estimated value of psychological inflexibility (M = 3.51, SD = 2.74) required for 50% of the sample to score equal to or above the increment X on the scale.

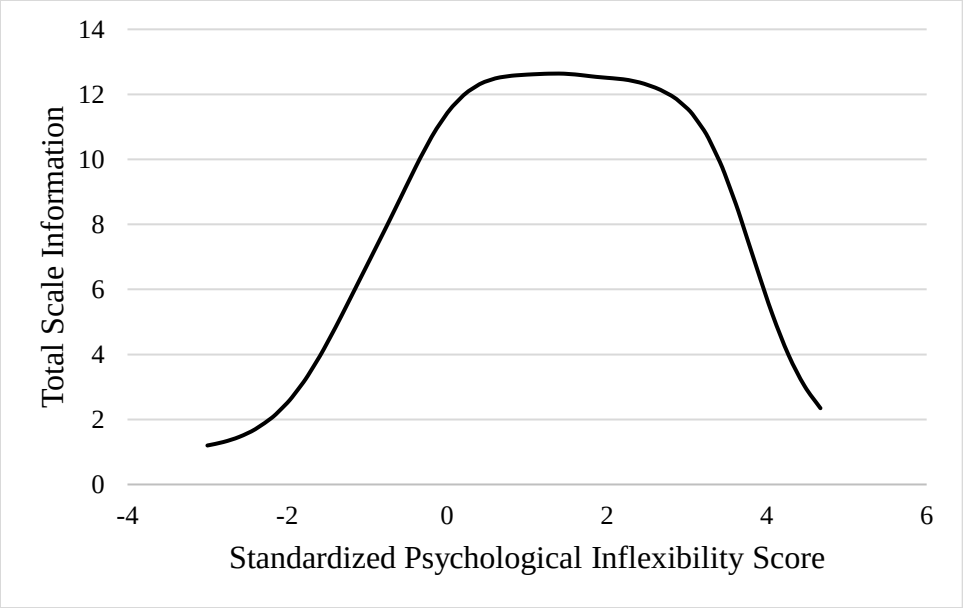

Figure 1. Item information function for all AAQ-II items across analysis groups.

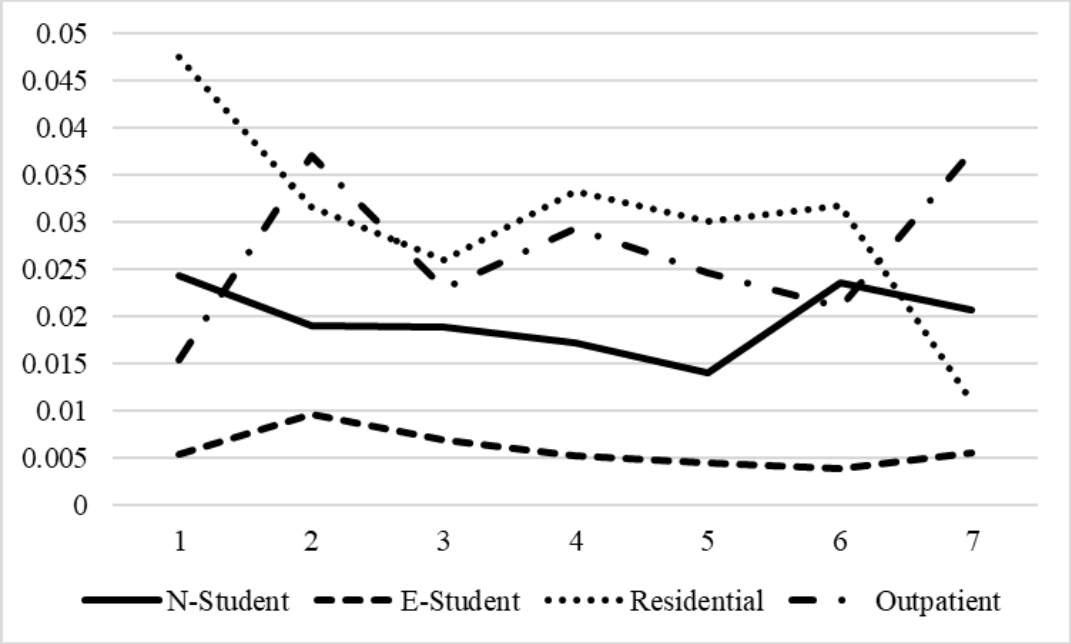

Figure 2. Mean absolute residual associated with all AAQ-II items, averaged across response categories.

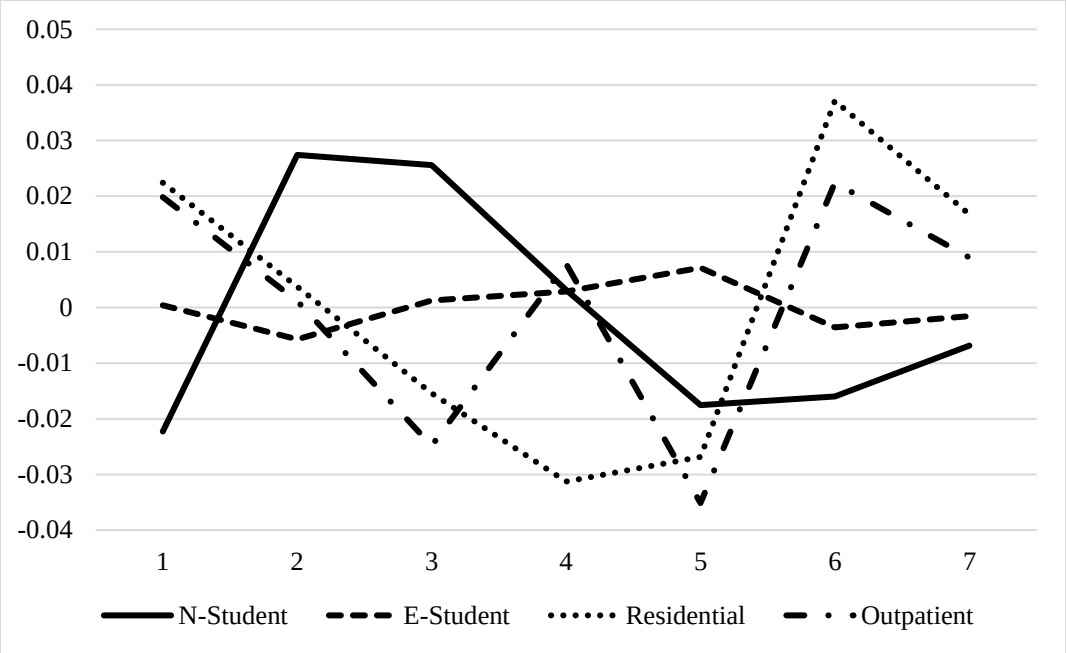

Figure 3. Mean residual associated with response categories 1-7, averaged across items.