Measuring Art Knowledge:
Item Response Theory and Differential Item Functioning
Analysis of the Aesthetic Fluency Scale

Katherine N. Cotter[1], David F. Chen[2], Alexander P. Christensen[1],

Kyung Yong Kim[2], and Paul J. Silvia[2]

[1]University of Pennsylvania

[2]University of North Carolina at Greensboro

AUTHOR NOTE

Katherine N. Cotter, Positive Psychology Center, University of Pennsylvania; David F. Chen and Kyung Yong Kim, Department of Educational Research Methodology, University of North Carolina at Greensboro; Alexander P. Christensen, Penn Center for Neuroaesthetics, Department of Neurology, University of Pennsylvania; Paul J. Silvia, Department of Psychology, University of North Carolina at Greensboro.

Thank you to Jeff Smith, Lisa Smith, and Eva Specker for sharing data for these analyses.

Correspondence regarding this article should be addressed to Katherine N. Cotter, Positive Psychology Center, University of Pennsylvania, Philadelphia, PA; email: katherinencotter@gmail.com.

Abstract

The Aesthetic Fluency Scale is a commonly used measure of people's art knowledge. This scale was initially developed for museum visitors, but its usage has expanded to other populations, including non-arts students. The present research used an Item Response Theory approach to better understand the scale's functioning in two samples—artistically engaged individuals (i.e., museum visitors and art students) and non-arts students—and any differences in scale properties between the samples (i.e., differential item functioning). Overall, terms related to art styles were easiest, the non-arts students had lower scores than the artistically engaged, and most items showed marked differences between the two samples. These results suggest that using this scale to draw comparisons between these populations is inappropriate. Our results also identify avenues for future development of the scale, including expanding the pool of terms used and revisiting the number of response options.

Keywords: art; aesthetics; aesthetic fluency scale; assessment; art education; item response theory

Measuring Art Knowledge:

Item Response Theory and Differential Item Functioning

Analysis of the Aesthetic Fluency Scale

People's knowledge of the arts uniquely shapes how they understand and interact with art and has been measured in a multitude of ways. Researchers have approached measurement of art expertise via training in art or art history (e.g., Belke et al., 2010; Chamberlain & Wagermans, 2015; Leder et al., 2014; Wanzer et al., 2020), behaviors related to art engagement (e.g., visiting museums; Chatterjee et al., 2010; Specker et al., 2020; Wanzer et al., 2020), performance on tests of art knowledge (Specker et al., 2020), and self-reported art expertise (Pelowski, 2015; Smith & Smith, 2006). In many studies, the assessment of art expertise is done ad hoc (e.g., Belke et al., 2010; Chamberlain & Wagermans, 2015; Leder et al., 2014; Pelowski, 2015), but there are a few scales that have been developed for the assessment of art expertise (i.e., Chatterjee et al., 2010; Smith & Smith, 2006; Specker et al., 2020). In the present research we focus on the Aesthetic Fluency Scale (Smith & Smith, 2006).

Aesthetic fluency is described as "the knowledge base concerning art that facilitates aesthetic experience in individuals" (Smith & Smith, 2006, p. 50). The Aesthetic Fluency Scale (Smith & Smith, 2006), one of the most prominent scales for measuring art knowledge, was developed to measure people's art knowledge via assessing self-reported understanding and knowledge of artists and art styles, periods, and concepts. For each of 10 items, people judge—from never knowing the artist or art idea existed to possessing detailed knowledge of the artist or art idea—where their expertise lies.

In their initial investigation of the scale with visitors to the Metropolitan Museum of Art, Smith and Smith (2006) examined the factor structure of the scale and how self-reported

aesthetic fluency was associated with the demographics of visitors. Of the ten items, visitors reported the greatest knowledge of Impressionism and the least knowledge of Chinese Scrolls. Interestingly, for many items in the scale, visitors indicated only a vague knowledge of the items, suggesting that the scale was challenging for many museum visitors. Principal components analysis supported retaining a single factor, which showed good internal consistency ($\alpha = .90$). Aesthetic fluency scores were positively predicted by visitor age, frequency of art museum visitation, and amount of training in art or art history; aesthetic fluency scores were not predicted by overall level of education. This pattern of findings is consistent with Smith and Smith's (2006) conceptualization of aesthetic fluency—knowledge of art is accumulated through multiple means, including formal instruction (e.g., studying art or art history) and informal instruction (e.g., visiting museums), and as age increases, people have had more time to accumulate this fluency.

Since its introduction, the Aesthetic Fluency Scale has continued to be used to examine people's knowledge of visual art. Some of this work has used this scale with museum visitors, art experts, or art students (Atari et al., 2020; Mullennix & Robinet, 2018; Smith & Smith, 2006, Specker et al., 2017; Specker et al., 2020). Beyond Smith and Smith's (2006) initial study developing and providing preliminary validation for the scale, however, the Aesthetic Fluency Scale has not undergone formal psychometric validation efforts with museum visitor or art expert populations, and researchers are increasingly using this measure to assess art knowledge in general undergraduate student samples (DeWall et al., 2011; Donato, 2019; Fayn et al., 2015; Fayn et al., 2018; Harrison & Clark, 2016; McKibben & Silvia, 2015; Nusbaum & Silvia, 2011; Silvia, 2007, 2013; Silvia & Barona, 2009; Silvia et al., 2009; Silvia & Nusbaum, 2011).

Within the general student samples, the scale has typically shown acceptable reliability ($\alpha > .80$; DeWall et al., 2011; Fayn et al., 2015; Fayn et al., 2018; Nusbaum & Silvia, 2011, 2013; Silvia, 2007; Silvia & Barona, 2009; Silvia & Nusbaum, 2011), but students generally indicate very low levels of familiarity with the terms (e.g., McKibben & Silvia, 2015; Silvia, 2007; Silvia & Barona, 2009). Across all projects, however, the results have been theoretically consistent—art experts have higher aesthetic fluency than undergraduate students (Atari et al., 2020; Mullennix et al., 2018), and people higher in aesthetic fluency spend more time engaged with the arts (Atari et al., 2020; Smith & Smith, 2006), report higher openness to experience (Atari et al., 2020; Fayn et al., 2015; Silvia, 2007), and are older (i.e., have had greater time to accumulate their aesthetic fluency; Atari et al., 2020; Smith & Smith, 2006).

But the Aesthetic Fluency Scale is not the only measure that has been used to assess art expertise—Chatterjee et al.'s (2010) Art Experience Questionnaire and Specker et al.'s (2020) Vienna Art Interest and Art Knowledge (VAIAK) questionnaire have also been widely used in the measurement of art expertise. The Art Experience Questionnaire (Chatterjee et al., 2010) contains eight self-report items—three items about formal education, two items about museum and gallery visitation, and three items regarding art-related engagement (i.e., art making, reading about art, looking at art). Each person's art expertise is determined through a composite score of these items. The VAIAK (Specker et al., 2020) breaks art expertise down into two components—art knowledge and art interest. Art knowledge is assessed through six multiple choice questions and 10 items in which people must identify the artist and style of the presented artworks (26 items in total). Art knowledge scores represent the number of items answered correctly. Art interest is assessed through 11 self-report items about frequency of art-related engagement (e.g., attending talks about art, reading about art) and general interest in art. The VAIAK yields

separate scores for the two components of art expertise. Like the Aesthetic Fluency Scale, both the Art Experience Questionnaire (e.g., Bromberger et al., 2011; Chatterjee et al., 2010; Ticini et al., 2014; van Paasschen et al., 2015) and the VAIAK (e.g., Gartus et al., 2020; Grüner et al., in press; Specker et al., 2020; Steciuch et al., in press) have been used to assess art expertise in student and expert samples.

The Art Experience Questionnaire, VAIAK, and the Aesthetic Fluency Scale each take a unique approach to the measurement of art expertise. The Art Experience Questionnaire emphasizes specific behaviors (i.e., classes taken, art viewing) that blends accumulation of art knowledge and interest, an approach that has its limitations (see Specker et al., 2020). The VAIAK separates these two components, defining art interest by both specific behaviors and attitudes indicating interest in the arts and assessing art knowledge using objective items regarding artist, style, and iconography identification. The Aesthetic Fluency Scale focuses on just the art knowledge component of expertise (though the scale is correlated with indices of art interest, e.g., Smith & Smith, 2006) focusing on self-reported knowledge of artists, styles, and art concepts. Supporting its validity, the Aesthetic Fluency Scale is associated with the objective assessment of knowledge via the VAIAK in a mixed student and expert sample (r = .62, Specker et al., 2020), suggesting that a self-report approach can be one way to examine art knowledge. Our interest in the Aesthetic Fluency Scale lies in its widespread usage with both novice and expert samples and its innovation in thinking about how art expertise can be assessed. This scale was one of the first systematic, scale-based methods of assessing art expertise to move beyond defining expertise by prior art education or specific art engagement behaviors. Further evaluation of this scale will provide additional insight as to its properties and suggestions for future usage.

In the present research, we used an Item Response Theory (IRT) approach to better understand the Aesthetic Fluency Scale's properties. An IRT approach treats scale items as the unit of analysis and permits the examination of specific qualities of the items, such as how hard each is or how well each distinguishes between people with differing levels of the construct measured by the scale. Our examination of the Aesthetic Fluency Scale using an IRT approach is one of the first to focus on how the scale functions in artistically engaged and non-arts student samples. This scale was originally developed with artistically engaged individuals in mind (e.g., museum goers; Smith & Smith, 2006), yet it is commonly used with non-arts student samples. Therefore, it is important to evaluate how the scale functions within these samples to better understand whether the scale captures meaningful differentiation in these samples. We expect that people with greater experience with art (i.e., the artistically engaged, such as museumgoers or art students) to indicate greater art knowledge than students not studying art. Moreover, it is important to understand whether the scale has the same meaning to both groups—that is whether the Aesthetic Fluency Scale is measuring aesthetic fluency similarly in non-arts student and artistically engaged samples (Putnick & Bornstein, 2016).

There were two primary aims for this investigation. The first was to better understand the difficulty of the scale and its items within the two groups and to determine whether any items are too easy or too difficult within the groups. The second aim was to examine measurement invariance between the two groups (i.e., is there evidence of differential item functioning), that is whether the items function similarly in the two groups. As this project was largely exploratory, we do not have hypotheses regarding whether items would show adequate difficulty or display measurement invariance across the two groups.

Method

Participants

Participants were 3,233 people who completed the Aesthetic Fluency Scale. The data were compiled from multiple projects involving undergraduate students with a variety of non-arts majors, art and art history students, and art museum visitors (see Table 1 for descriptions of the individual samples and demographic information). This sample was divided into two groups—undergraduate students with majors unrelated to the visual arts (n = 1,791) and artistically engaged individuals (i.e., students studying visual art, art museum visitors; n = 1,442)[1]. The purpose of using these subsamples was to examine the properties of the scale within the population the scale was developed for use with (i.e., artistically engaged individuals) and the properties of the scale in an increasingly studied population (i.e., undergraduate students with majors unrelated to the visual arts).

Aesthetic Fluency Scale

All participants completed the Aesthetic Fluency Scale (Smith & Smith, 2006). This scale contains 10 items referring to different artists (i.e., Mary Cassatt, Isamu Noguchi, John Singer Sargent, Alessandro Botticelli, Gian Lorenzo Bernini), artistic styles (i.e., Impressionism, Fauvism, Abstract Expressionism), or art concepts (i.e., Egyptian Funerary Stelae, Chinese Scrolls). Participants indicated their familiarity with each of the items on a 5-point Likert scale (0 = “I have never heard of this artist or term”; 1 = “I have heard of this but don’t really know anything about it”; 2 = “I have a vague idea of what this is”; 3 = “I understand this artist or idea

[1]Museum visitors and students studying visual art were grouped together because there were not enough art students to examine them as a separate group. The arts students were grouped with the museum visitors (rather than with the non-arts students) due to their engagement in visual art-related activities.

when it is discussed"; 4 = "I can talk intelligently about this artist or idea in art"). The scale score is calculated by summing responses to all items, so the possible scale scores range from 0 to 40.

Graded Response Model

We used a unidimensional graded response model to estimate parameters—this is the first step in addressing our aim to evaluate whether items and the test displayed adequate difficulty. The graded response model (GRM; Samejima, 1969) handles data with rankings or ordered polytomous categories, such as responses to attitudinal assessments, and is appropriate for use with Likert-type response scales (Lautenschlager, et al., 2006; Robie et al., 2001). From this model, we can obtain two types of estimates for each item: *discrimination parameters* and *boundary location parameters*. The discrimination parameter (*a*) indicates how strongly the item is associated with the latent variable being assessed. This can be seen in item category response functions (ICRFs) as lines with steeper slopes. Each item has its own discrimination parameter. The item boundary location parameters indicate the ability level (θ)—here, the level of aesthetic fluency—at which the probability of indicating a higher level of ability is 0.50. The number of item boundary parameters estimated depends on the number of response options. For the Aesthetic Fluency Scale, there are four item boundary location parameters (b_1-b_4) estimated for each item because there are five response options. For example, b_1 represents the ability level at which the probability of responding with a 1 (i.e., "I have heard of this but don't really know anything about it") or greater is equal to 0.50, b_2 represents the ability level at which the probability of responding with a 2 or greater is equal to 0.50, and so forth.

The GRM requires that the thresholds between response categories are ordered, which aligns with the structure of the Aesthetic Fluency Scale's response options—someone who claims the ability to talk intelligently about an artist must also be familiar with that artist and

understand when the artist is discussed. The GRM allows for free estimation of all item parameters (i.e., not constraining parameters to be equal across items). Although this results in a more complex model, the present study was interested in understanding the variability in difficulty and discrimination among the scale items. For these reasons we elected to use a GRM over alternate polytomous IRT approaches.

As θ is a latent variable in IRT models, it should be assigned a scale. Typically, the θ-scale is set by assuming that the group is a sample from a population that is normally distributed with a mean of 0 and a standard deviation of 1. Therefore, a θ of 0 represents average ability and a θ of 1, for example, indicates being one standard deviation above average. This scaling procedure was used for the present analyses.

Additionally, we can obtain the *information* for each item and the complete scale in the form of item and test information curves. Item and test information are used to determine how much information each item and the entire scale provide for the estimation of ability along the entire scale. The peak of these curves indicates the ability level for which the item is most informative. For the present project, information tells us how much information each item and the scale as a whole provide for the estimation of aesthetic fluency levels.

Linking Parameter Estimates

Two sets of item parameter estimates were obtained for each item—one from the artistically engaged sample and one from non-arts student sample. These item estimates must be linked together before they can be appropriately compared. Linking is a statistical procedure that adjusts for differences in group ability by placing the item parameter estimates from one group onto the scale of another group. One method for linking the item estimates, used in this analysis, is to calibrate the two groups' responses simultaneously in one computer run. This method,

referred to as concurrent calibration, estimates the groups' distributions simultaneously with the item parameters. Thus, the estimated distributions and item parameter estimates obtained from concurrent calibration were already on the same scale. In the present analysis, all parameter estimates are on the artistically engaged sample's scale.

Differential Item Functioning

After linking, it is expected that the item parameter estimates for each group of examinees will be similar for each item due to the IRT property of parameter invariance. That is, the parameters should remain equal across groups of respondents and measurement conditions. For example, an artistically engaged visitor with a θ of 0 should have the same probability of obtaining a score of 1 or greater to an item as non-arts student with a θ of 0. If the probabilities differ, then there is evidence of differential item functioning.

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014), differential item functioning is an issue of fairness because the probability of correctly answering an item is based upon group membership, as opposed to ability. Differential item functioning represents a source of construct-irrelevant variance—extraneous factors influencing scale scores—that compromises the valid interpretation of scores and the ensuing decisions made from those scores. When items exhibit differential item functioning, they are often removed from the scale or test.

Results

Because we are using a unidimensional IRT model, we first performed exploratory factor analyses using the correlation matrix for the two samples to confirm that the Aesthetic Fluency Scale measures a single dimension. Analyses were completed using the *psych* package (Revelle, 2020) in R (R Core Team, 2020). The scree plots (see Figure 1) for each sample suggest

retaining one dimension; similarly, if using a cutoff criterion of eigenvalues greater than one, analyses for both samples also indicate retaining one dimension. Coefficient omega, an index of unidimensionality appropriate for use with ordinal scales (Peters, 2014), was acceptable for the artistically engaged ($\Omega = .90$) and non-arts student subsamples ($\Omega = .76$). These findings suggest that proceeding with unidimensional analyses is appropriate, in line with Smith and Smith's (2006) assessment of dimensionality.

A unidimensional GRM was used to calibrate the discrimination and boundary location parameters for the 10 items. All calibrations and differential item functioning analyses were completed using the *mirt* package (Chalmers, 2012) in R. We used a multiple group analysis with concurrent calibration to obtain item parameter estimates for both groups—the artistically engaged sample was used as the reference group, and the estimates for the non-arts student sample were placed on the reference group scale (i.e., a θ of 0 indicates average aesthetic fluency for the artistically engaged sample for both sets of estimates). As part of this calibration, we specified the Gian Lorenzo Bernini item as an anchor item (i.e., fixing the estimates to be equal across groups) in order to calibrate the remaining items. This item was selected using a Wald 2 test to identify items that are unlikely to exhibit differential item functioning, a method suggested by Woods et al. (2012) to determine appropriate anchor items. For the differential item functioning analyses, we used the likelihood ratio test with the forward method as it was possible that many items would show differential item functioning.

Descriptive statistics for all items are available in Table 2. Parameter estimates for the artistically engaged individuals in Table 3 and for non-arts students in Table 4, and Table 5 provides the differential item functioning results. Figures 2-5 display the ICRFs (Figures 2-3) for

the artistically engaged and non-arts student samples, respectively, and item information (Figures 4) and test information (Figure 5) for both samples.

Artistically Engaged Sample

Participants in the artistically engaged sample had an average aesthetic fluency score of 15.92 (*SD* = 9.16, *range* = 0 – 40) and a latent mean of 0 (*SD* = 1)[2]. People indicated the greatest amount of knowledge of terms related to art styles and similar amounts of knowledge for the artist and art concept terms. Specifically, people endorsed the highest knowledge of Impressionism (*M* = 2.87, *SD* = 1.04) and Abstract Expressionism (*M* = 2.19, *SD* = 1.24) and the least knowledge of Isamu Noguchi (*M* = .94, *SD* = 1.30).

Item discrimination parameters ranged from 0.97 to 2.60—the least discriminating item was Chinese Scrolls (*a* = 0.97), and the most discriminating item was Mary Cassatt (*a* = 2.60; see Table 3). Collectively, artist items were the most discriminating item type and art concept items were the least discriminating item type.

Within the artistically engaged sample, the items demonstrated a range of difficulty (see Table 3 and Figure 2). For all items (except Isamu Noguchi), the first boundary parameters (b_1) were negative, indicating that people with below average aesthetic fluency were able to report some level of knowledge of the terms (i.e., respond with a 1 or higher). Of the ten items, Impressionism (b_1 = -2.26) and Abstract Expressionism (b_1 = -1.65) required the lowest aesthetic fluency level to endorse knowledge of the term. Indicating greater than a slight familiarity with a term (i.e., reporting having at least a vague idea of what the term is) required average or slightly above average ability for most items, however. Endorsing a high level of knowledge (i.e.,

[2] Because the artistically engaged sample was used as the reference group, the latent mean and standard deviation were fixed to 1 and 0, respectively.

reporting the ability to talk intelligently about the term) generally required an aesthetic fluency level over one standard deviation above average. Impressionism was the only term with a final boundary less than 1 (b_4 = 0.59); the two art concept terms—Egyptian Funerary Stelae (b_4 = 2.97) and Chinese Scrolls (b_4 = 3.89)—were the only items to have parameters greater than 2, suggesting it would be unlikely that someone not 3 standard deviations above average in their aesthetic fluency would indicate a high level of familiarity with these terms.

The test information (see Figure 5) suggests the aesthetic fluency scale provides the most information for people near average aesthetic fluency and little information for those at extremely low (i.e., $\theta < -2$) or high (i.e., $\theta > 2$) levels of aesthetic fluency. But the individual items showed variability in information provided (see Figure 4). The Egyptian Funerary Stelae and Chinese Scrolls items provided minimal information at all levels of aesthetic fluency. The artist items were most informative for people with near average aesthetic fluency, with Mary Cassatt being the most informative artist item. For the art style items, Impressionism and Abstract Expressionism provided the most information for people with low aesthetic fluency and Fauvism provided the most information for people with near average aesthetic fluency.

Non-Arts Student Sample

Non-arts students had an average aesthetic fluency score equal to 7.13 (*SD* = 5.42, *range* = 0 – 40) and a latent mean of -1.35 (*SD* = .57), indicating that the non-arts students reported lower levels of and variation in aesthetic fluency than the artistically engaged sample. Like the artistically engaged sample, people indicated greater knowledge of art style terms and less knowledge of artist and art concept terms. Students indicated the most knowledge regarding Impressionism (*M* = 1.77, *SD* = 1.12) and Abstract Expressionism (*M* = 1.54, *SD* = 1.16) and the least amount of knowledge regarding Isamu Noguchi (*M* = .10, *SD* = .42). In fact, the median

response for all but three terms (Impressionism, Chinese Scrolls, and Abstract Expressionism) was 0, meaning most students reported not having heard of most of the terms, unlike the artistically engaged individuals who only had a median response of 0 for one item (Isamu Noguchi).

Differential Item Functioning

We conducted differential item functioning analyses on nine items (Gian Lorenzo Bernini served as an anchor item) to determine whether the items function similarly in the artistically engaged and non-arts student samples. All but one of the test items (Egyptian Funerary Stelae) showed signs of differential item functioning, suggesting that the Aesthetic Fluency Scale largely does not operate similarly in the two populations examined.

When examining the item parameter estimates (Tables 3 and 4) and item information curves (Figure 4), there are a few interesting differences between the samples. The Impressionism and Abstract Expressionism items were more discriminating for non-arts students and more informative for non-arts students than for artistically engaged individuals, and the John Singer Sargent and Alessandro Botticelli items were more informative for artistically engaged individuals than for non-arts students. The John Singer Sargent item also was more difficult for non-arts students, especially at higher levels of familiarity (indicated by $b_{2,}$ $b_{3,}$ and b_4), whereas the Chinese Scrolls item was more difficult for artistically engaged individuals at all familiarity levels. The test information curves for both samples are generally similar—both show the greatest information for people near $\theta = 0$ levels of aesthetic fluency with information sharply decreasing for people with high ability (i.e., $\theta > 2$)—but the test information for non-arts student sample shows greater information at below $\theta = 0$ levels of fluency than the test information for the artistically engaged sample.

Discussion

The present research used an IRT approach to examine the Aesthetic Fluency Scale (Smith & Smith, 2006) in samples of artistically engaged individuals and non-arts students using an IRT approach. The results suggest that the items vary in their difficulty, with items related to art styles being easiest and art concept and artist items being more difficult, and how well they discriminate between people with low and high levels of aesthetic fluency. The scale appears to be most useful for people who have aesthetic fluency levels near the artistically engaged sample's mean. But most of the items showed differential item functioning, suggesting that the scale operates differently in artistically engaged and non-arts student samples and that making comparisons between these two groups using this scale is inappropriate.

Unsurprisingly, the artistically engaged individuals had higher aesthetic fluency than did the non-arts students. This pattern is largely to be expected—a component of most art education programs involves art history coursework, increasing the likelihood of exposure to the terms, and museumgoers may engage in informal learning through exposure to different artworks, artists, and museum programs during their visits (Smith & Smith, 2006). Additionally, it is important to recognize that because the Aesthetic Fluency Scale is a measure of self-reported knowledge, it may be susceptible to social desirability. Because members of the artistically engaged sample may feel that they are expected to have higher levels of art knowledge, they may report greater knowledge than they possess. This same pressure may not exist for non-arts students and so group differences may, in part, be due to this factor. It should be noted, however, that prior work has not found responses to this scale to be associated with social desirability in a student sample and that when a non-existent art term was added to the scale, few respondents claimed knowledge of this item (McKibben & Silvia, 2015).

These analyses have several implications for future use of the Aesthetic Fluency Scale. First, there are implications for the scale's usage in different populations. There was a large difference between the two samples examined here—the artistically engaged sample had a higher aesthetic fluency latent mean than the non-arts students and most items exhibited differential item functioning. The non-arts students appeared to struggle with the difficulty of the scale—for seven of the ten items, the median response was never having heard of the term, the lowest possible option. Six percent of the non-arts students reported never hearing of any of the terms, and over 75% of the sample had an overall score less than 10 out of 40; only 2% of the artistically engaged sample reported never hearing of any terms and 29% of the artistically engaged had an overall score less than 10.

This pattern suggests that the Aesthetic Fluency Scale in its current form suffers from a floor effect with non-arts students, and researchers interested in measuring art knowledge may consider using a different scale in this population (e.g., the Vienna Art Interest and Art Knowledge scale; Specker et al., 2020). This also presents the opportunity for additional development of the Aesthetic Fluency Scale to include additional art style items (as these were the easiest item type) or names of artists more widely known (e.g., Monet, Picasso) to better measure aesthetic fluency in non-arts students. Researchers who are using this scale in its current form with non-arts students may consider transforming people's scores to reduce the influence of the students who do report high levels of aesthetic fluency and account for the positively skewed distribution (Bland et al., 2013; Peterson & Cavanaugh, 2020). Although these findings do not indicate the Aesthetic Fluency Scale cannot be used in non-art student populations, the difficulty of the scale and consistently low scores of non-arts students may make it difficult to identify meaningful differences among people with lower levels of art knowledge. Measures that are

better equipped to address this range of art knowledge may be better suited for use in the general student population.

Additionally, most items showed signs of differential item functioning, suggesting that the scale does not operate similarly in the two populations examined. This poses a threat to the validity of the interpretation and use of scores (AERA et al., 2014) and suggests that making comparisons between groups using traditional scoring procedures (i.e., sum scores) that do not account for differential item functioning is inadvisable. Alternatively, some of the differences between groups may be attributable to the considerable differences in the composition of the two samples (e.g., age differences; see Table 1) or differences in testing conditions between the groups—the non-arts student samples were typically tested as part of a course research option, often in university labs, whereas the artistically engaged sample were typically tested during the course of their visit to an art museum.

Second, IRT analyses allow us to evaluate individual items. One method of examining items is looking at information each item provides about someone's aesthetic fluency levels (see Figure 4). Different items will often be more informative at different levels of aesthetic fluency (i.e., some items will tell us more about people with near average aesthetic fluency whereas others may be more informative for people with somewhat above average aesthetic fluency). In this study, however, the Egyptian Funerary Stelae item provided virtually no information at all levels of aesthetic fluency in both groups. Because a person's response to this item does not provide substantial information about their aesthetic fluency level, removing this item may improve the performance of the scale (e.g., Lalor et al., 2016). Similarly, the Chinese Scrolls item was much more difficult for the artistically engaged sample than the non-art student sample, suggesting that this item may not be interpreted in the same way by these samples—that is, non-

art student samples might be confusing this item with something different than is intended. This is further evidence that additional development on this scale may be beneficial.

Finally, IRT analyses allow us to examine the response options of the scale using the ICRFs (see Figures 2 and 3). The trace lines in these figures show the probability of giving a particular response to an item at a particular level of aesthetic fluency. In cases where the trace lines completely overlap, this suggests that one response is never the most likely response at any given aesthetic fluency level. Across the full continuum of aesthetic fluency levels, it appears that only three responses consistently have trace lines not fully overlapping with other lines (i.e., have aesthetic fluency levels when that response is the most likely). This may suggest that this scale could benefit from reducing the number of response options, perhaps from five to three, to more effectively assess aesthetic fluency (e.g., Linacre, 2002; Muraki, 1993). This decrease in scale response options would likely result in little information being lost but may prove beneficial for estimation purposes, increasing model parsimony, and faster administration of the scale.[3]

The Aesthetic Fluency Scale has been widely used to measure people's art expertise. Overall, our findings suggest this scale may be best suited for artistically engaged individuals, such as art museum visitors or students studying visual art. The brevity of the Aesthetic Fluency Scale makes it a museum-friendly tool for visitors with limited time (and patience) for research participation, but we often have fewer time-constraints with student samples, allowing us to use more comprehensive measures of art expertise. Our findings also provide avenues for additional development of the Aesthetic Fluency Scale that may make it more student-friendly and improve the precision of its measurement.

[3] We thank an anonymous reviewer who raised this point.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Atari, M., Afhami, R., Mohammadi-Zarghan, S. (2020). Exploring aesthetic fluency: The roles of personality, nature relatedness, and art activities. *Psychology of Aesthetics, Creativity, and the Arts, 14*(1), 125-131.

Belke, B., Leder, H., Harsanyi, G., & Carbon, C. C. (2010). When a Picasso is a "Picasso": The entry point in the identification of visual art. *Acta Psychologica, 133,*191-202.

Bland, J. M., Altman, D. G., & Rohlf, F. J. (2013). In defence of logarithmic transformations. *Statistics in Medicine, 32,* 3766-3769.

Bromberger, B., Sternschein, R., Widick, P., Smith, W., II, & Chatterjee, A. (2011). The right hemisphere in esthetic perception. *Frontiers in Human Neuroscience, 5*, 109.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chamberlain, R. & Wagemans, J. (2015). Visual arts training is linked to flexible attention to local and global levels of visual stimuli. *Acta Psychologica, 161,* 185-197.

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., II, & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts, 28,* 207-222.

Cotter, K. N., Silvia, P. J., Bertamini, M., Palumbo, L., & Vartanian, O. (2017). Curve appeal: Exploring individual differences in preference for curved versus angular objects. *i-Perception,* March-April, 1-17.

Donato, F. V. (2019). *Visual ambiguity priming promotes uniqueness in art-viewing responses* [Unpublished doctoral dissertation]. Texas Tech University.

Fayn, K., Silvia, P. J., Erbas, Y., Tiliopoulos, N., & Kuppens, P. (2018). Nuanced aesthetic emotions: Emotion differentiation is related to knowledge of the arts and curiosity. *Cognition and Emotion, 32*(3), 593-599.

Gartus, A., Völker, M., & Leder, H. (2020). What experts appreciate in patterns: Art expertise modulates preference for asymmetric and face-like patterns. *Symmetry, 12*(5), 707.

Grüner, S., Specker, E., & Leder, H. (in press). Effects of context and genuineness in the experience of art. *Empirical Studies of the Arts.*

Lalor, J. P., Wu, H., & Yu, H. (2016). Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 648). NIH Public Access.

Lautenschlager, G. J., Meade, A. W., & Kim, S. H. (2006). Cautions regarding sample characteristics when using the graded response model. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, Texas.

Leder, H., Gerger, G., Brieber, D., & Schwarz, N. (2014). What makes an art expert? Emotion and evaluation in art appreciation. *Cognition and Emotion, 28,* 1137-1147.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Mullenix, J. W. & Robinet, J. (2018). Art expertise and the processing of titled abstract art. *Perception, 47*(4), 359-378.

Muraki, E. (1993). Information functions of the generalized partial credit model. *ETS Research Report Series*, *1993*(1), i-12.

Pelowski, M. (2015). Tears and transformation: Feeling like crying as an indicator of insightful or "aesthetic" experience with art. *Frontiers in Psychology*, *6*, 1006.

Peters, G. -J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist*, *16*(2), 56-69.

Peterson, R. A., & Cavanaugh, J. E. (2020). Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, *47*(13-15), 2312-2327.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, *14*(2), 187-207.

Rossell, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

Silvia, P. J. (2007). Knowledge-based assessment of expertise in the arts: Exploring aesthetic fluency. *Psychology of Aesthetics, Creativity, and the Arts, 1*(4), 247-249.

Silvia, P. J. (2013). Interested experts, confused novices: Art expertise and the knowledge emotions. *Empirical Studies of the Arts, 31*(1), 107-115.

Silvia, P. J. & Barona, C. M. (2009). Do people prefer curved objects? Angularity, expertise, and aesthetic preference. *Empirical Studies of the Arts, 27*(1), 25-42.

Smith, L. F. & Smith, J. K. (2006). The nature and growth of aesthetic fluency. In P. Locher, C. Martindale, & L. Dorfman (Eds.), *New directions in aesthetics, creativity, and the arts* (pp. 47-58). Amityville, NY: Baywood.

Specker, E., Forster, M., Brinkmann, H., Boddy, J., Pelowski, M., Rosenberg, R., & Leder, H. (2020). The Vienna Art Interest and Art Knowledge Questionnaire (VAIAK): A unified and validated measure of art interest and art knowledge. *Psychology of Aesthetics, Creativity, and the Arts, 14*(2), 172-185.

Steciuch, C. C., Kopatich, R. D., Feller, D. P., Durik, A. M., & Millis, K. (in press). Don't go with your gut: Exploring the role of motivation in aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts.*

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Ticini, L. F., Rachman, L., Pelletier, J., & Dubal, S. (2014). Enhancing aesthetic appreciation by priming canvases with actions that match the artist's painting style. *Frontiers in Human Neuroscience, 8*, 391.

van Paasschen, J., Bacci, F., & Melcher, D. P. (2015). The influence of art expertise and training on emotion and preference ratings for representational and abstract artworks. *PLoS One, 10*(8), e0134241.

Wanzer, D. L., Finley, K. P., Zarian, S., & Cortez, N. (2020). Experiencing flow while viewing art: Development of the aesthetic experience questionnaire. *Psychology of Aesthetics, Creativity, and the Arts, 14*(1), 113-124.

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1-33.

Table 1

Demographic Information for the Samples

	Collection Description	***N***			**Gender**	**Age**
		Total	**Artistically Engaged**	**Non-Arts Student**		
Sample 1	Participants were recruited as part of a course research participation option at a mid-sized Southeastern U.S. university	1,365	48	1,317	26.28% Male, 73.21% Female, 0.15% Unreported	*M* = 19.19 *SD* = 2.39 *Range* = 17 – 49
Sample 2	Participants were recruited during their visit to an art museum in New York City	960	960	0	31.20% Male, 55.22% Female, 13.58% Unreported	Under 18 = 42 18 – 24 = 96 25 – 34 = 124 35 – 44 = 125 45 – 54 = 158 55 – 64 = 169 65 – 74 = 135 75+ = 59
Sample 3	Participants were recruited for being psychology or art history students at a large Austrian university or for their expertise in art history	620	146	474	27.69% Male, 63.47% Female, 8.53% Unreported	*M* = 23.85 *SD* = 9.20 *Range* = 17 – 81
Sample 4	Participants were recruited during their visit to an art museum in the Southeastern U.S.	159	159	0	38.51% Male, 55.90% Female, 5.59% Unreported	*M* = 31.50 *SD* = 19.60 *Range* = 18 – 88
Sample 5	Participants were recruited during their visit to an art museum in the Netherlands	129	129	0	46.92% Male, 53.07% Female	*M* = 47.00 *SD* = 17.06 *Range* = 17 – 80

Non-Arts Student Sample	Includes undergraduate students who are not studying the visual arts	1,791	0	1,791	28.59% Male, 70.80% Female, 0.61% Unreported	*M* = 20.01 *SD* = 4.07 *Range* = 17 – 81
Artistically Engaged Sample	Includes art museum visitors and students studying the visual arts	1,442	1,442	0	32.04% Male, 58.74% Female, 9.92% Unreported	*M* = 43.62 *SD* = 18.54 *Range* = 17 – 88

Note. Because Sample 2 collected age information in age ranges, the midpoint for each range category was used in calculating the Artistically Engaged Sample age statistics.

Table 2

Descriptive statistics for Aesthetic Fluency Scale responses

	Full Sample	Artistically Engaged	Non-Arts Students
	M (*SD*), *Median*	*M* (*SD*), *Median*	*M* (*SD*), *Median*
Mary Cassatt	.78 (1.30) Med. = .00	1.55 (1.52) Med. = 1.00	.19 (.62) Med. = .00
Isamu Noguchi	.46 (1.00) Med. = .00	.94 (1.30) Med. = .00	.10 (.42) Med. = .00
John Singer Sargent	1.06 (1.29) Med. = 1.00	1.80 (1.46) Med. = 2.00	.50 (.75) Med. = .00
Alessandro Botticelli	1.17 (1.31) Med. = 1.00	1.90 (1.35) Med. = 2.00	.60 (.94) Med. = .00
Gian Lorenzo Bernini	.83 (1.21) Med. = .00	1.45 (1.37) Med. = 1.00	.37 (.79) Med. = .00
Fauvism	.81 (1.18) Med. = .00	1.34 (1.38) Med. = 1.00	.41 (.79) Med. = .00
Egyptian Funerary Stelae	.85 (1.12) Med. = .00	1.36 (1.23) Med. = 1.00	.46 (.83) Med. = .00
Impressionism	2.26 (1.21) Med. = 2.00	2.87 (1.04) Med. = 3.00	1.77 (1.12) Med. = 2.00
Chinese Scrolls	1.23 (1.09) Med. = 1.00	1.24 (1.16) Med. = 1.00	1.21 (1.03) Med. = 1.00
Abstract Expressionism	1.82 (1.24) Med. = 2.00	2.19 (1.24) Med. = 2.00	1.54 (1.16) Med. = 1.00
Total Scale Score	11.05 (8.53) Med. = 8.00	15.92 (9.16) Med. = 15.00	7.13 (5.42) Med. = 6.00

Table 3

Discrimination and boundary location parameter estimates for artistically-engaged individuals

	Discrimination (*a*)	Boundary 1 (b_1)	Boundary 2 (b_2)	Boundary 3 (b_3)	Boundary 4 (b_4)
Mary Cassatt	2.60	-0.31	0.01	0.44	1.25
Isamu Noguchi	2.12	0.18	0.66	1.18	1.88
John Singer Sargent	2.46	-0.73	-0.18	0.31	1.20
Alessandro Botticelli	2.49	-0.86	-0.40	0.29	1.41
Gian Lorenzo Bernini	1.92	-0.48	0.14	0.75	1.80
Fauvism	2.11	-0.32	0.24	0.79	1.73
Egyptian Funerary Stelae	1.13	-0.79	0.25	1.42	2.97
Impressionism	2.23	-2.26	-1.58	-0.66	0.59
Chinese Scrolls	0.97	-0.79	0.51	1.92	3.89
Abstract Expressionism	1.69	-1.65	-0.80	0.09	1.42

Table 4

Discrimination and boundary location parameter estimates for non-arts students

	Discrimination (*a*)	Boundary 1 (b_1)	Boundary 2 (b_2)	Boundary 3 (b_3)	Boundary 4 (b_4)
Mary Cassatt	2.33	-0.09	0.44	0.87	1.53
Isamu Noguchi	2.17	0.32	0.96	1.75	1.96
John Singer Sargent	0.74	-0.70	2.20	3.69	5.19
Alessandro Botticelli	1.61	-0.97	-0.07	0.63	1.91
Gian Lorenzo Bernini	1.92	-0.48	0.14	0.75	1.80
Fauvism	2.14	-0.69	0.02	0.74	1.59
Egyptian Funerary Stelae	1.29	-0.59	0.39	1.50	2.69
Impressionism	4.27	-2.22	-1.55	-0.85	-0.01
Chinese Scrolls	1.87	-1.98	-1.02	0.02	1.41
Abstract Expressionism	3.63	-2.00	-1.35	-0.69	0.12

Note. Estimates reported have been linked to the artistically engaged sample.

Table 5

Differential item functioning results

Item	χ^2	*p*
Mary Cassatt	$\chi^2(5) = 24.90$	< .001
Isamu Noguchi	$\chi^2(5) = 19.55$	.002
John Singer Sargent	$\chi^2(5) = 187.84$	< .001
Alessandro Botticelli	$\chi^2(5) = 50.03$	< .001
Fauvism	$\chi^2(5) = 8.97$	< .001
Egyptian Funerary Stelae	$\chi^2(5) = 8.97$	.110
Impressionism	$\chi^2(5) = 91.56$	< .001
Chinese Scrolls	$\chi^2(5) = 409.22$	< .001
Expressionism	$\chi^2(5) = 255.26$	< .001

Note. The Gian Lorenzo Bernini item served as an anchor item for these analyses and was not assessed for differential item functioning. The Bonferroni-correct significance level is .05/9 = .006.

Figure 1. Exploratory factor analysis scree plots

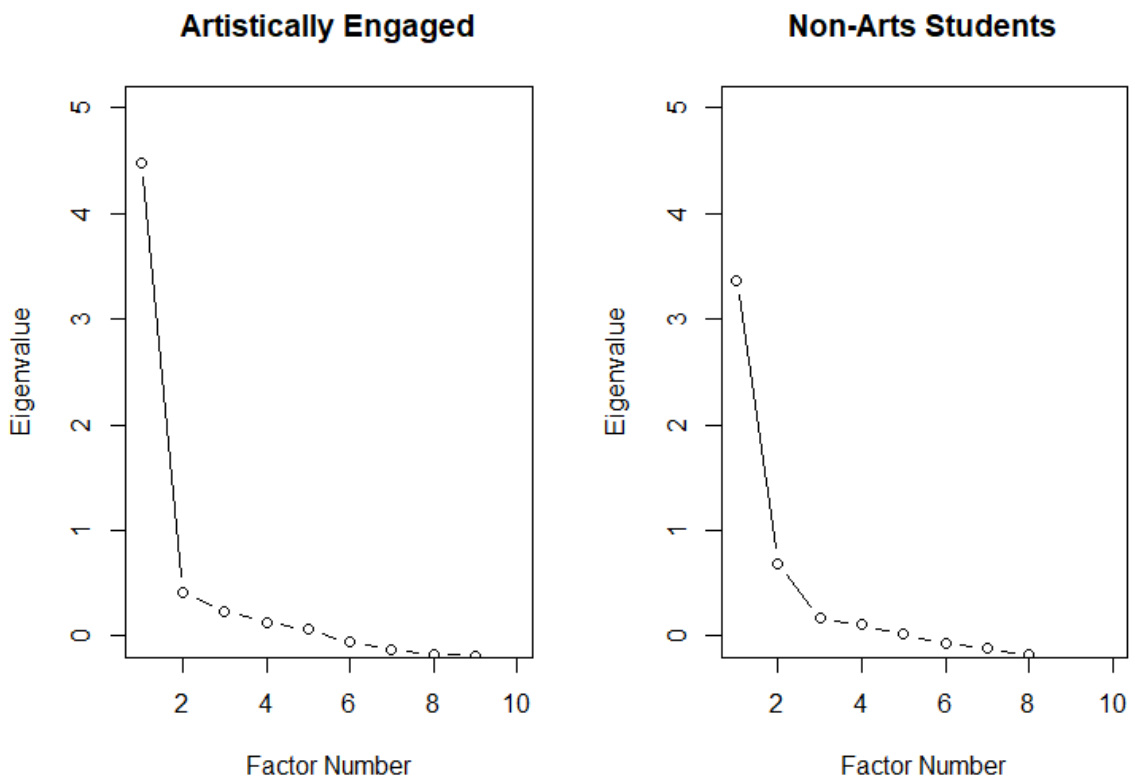

Figure 2. Item category response functions for the artistically engaged sample

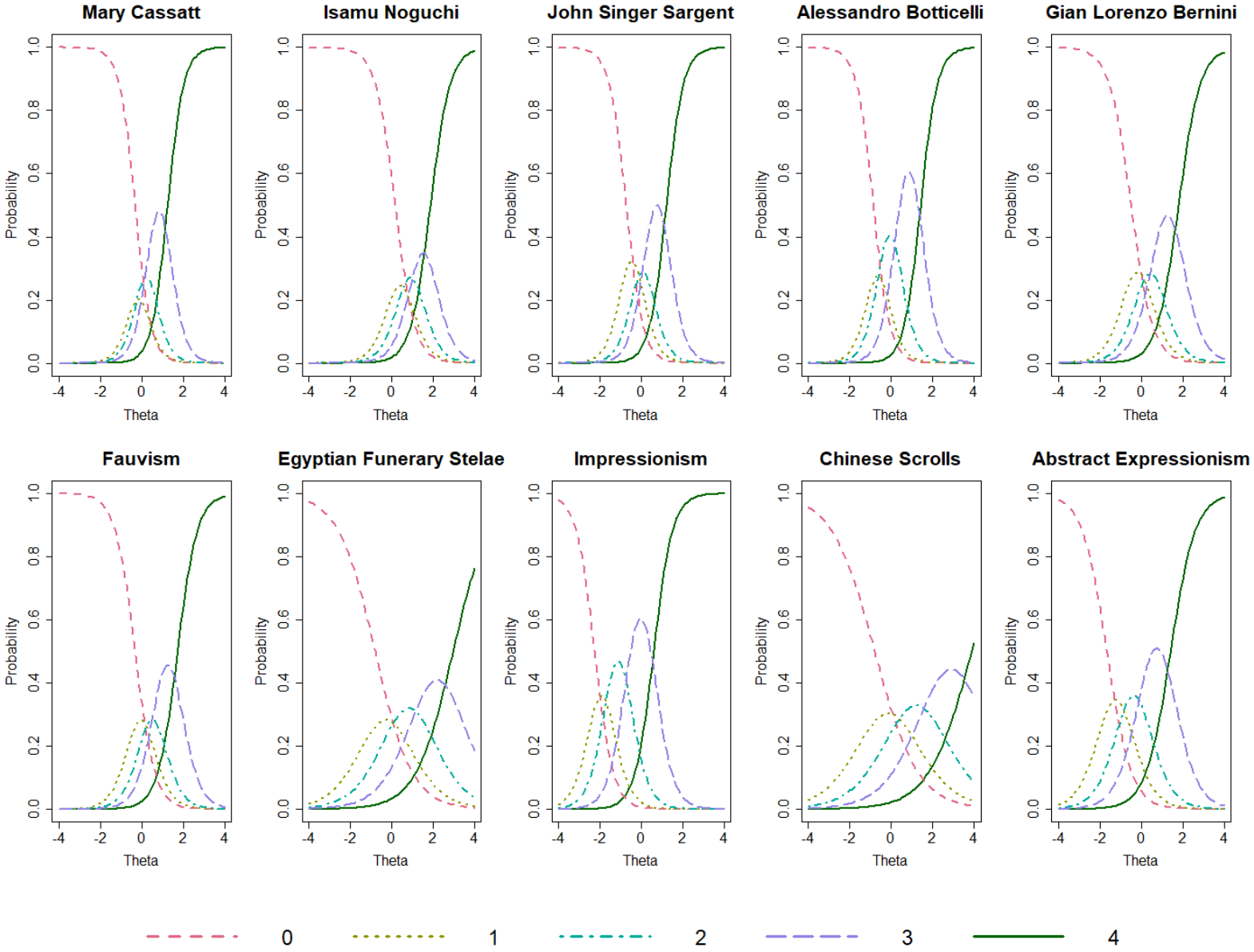

Figure 3. Item category response functions for the non-arts student sample

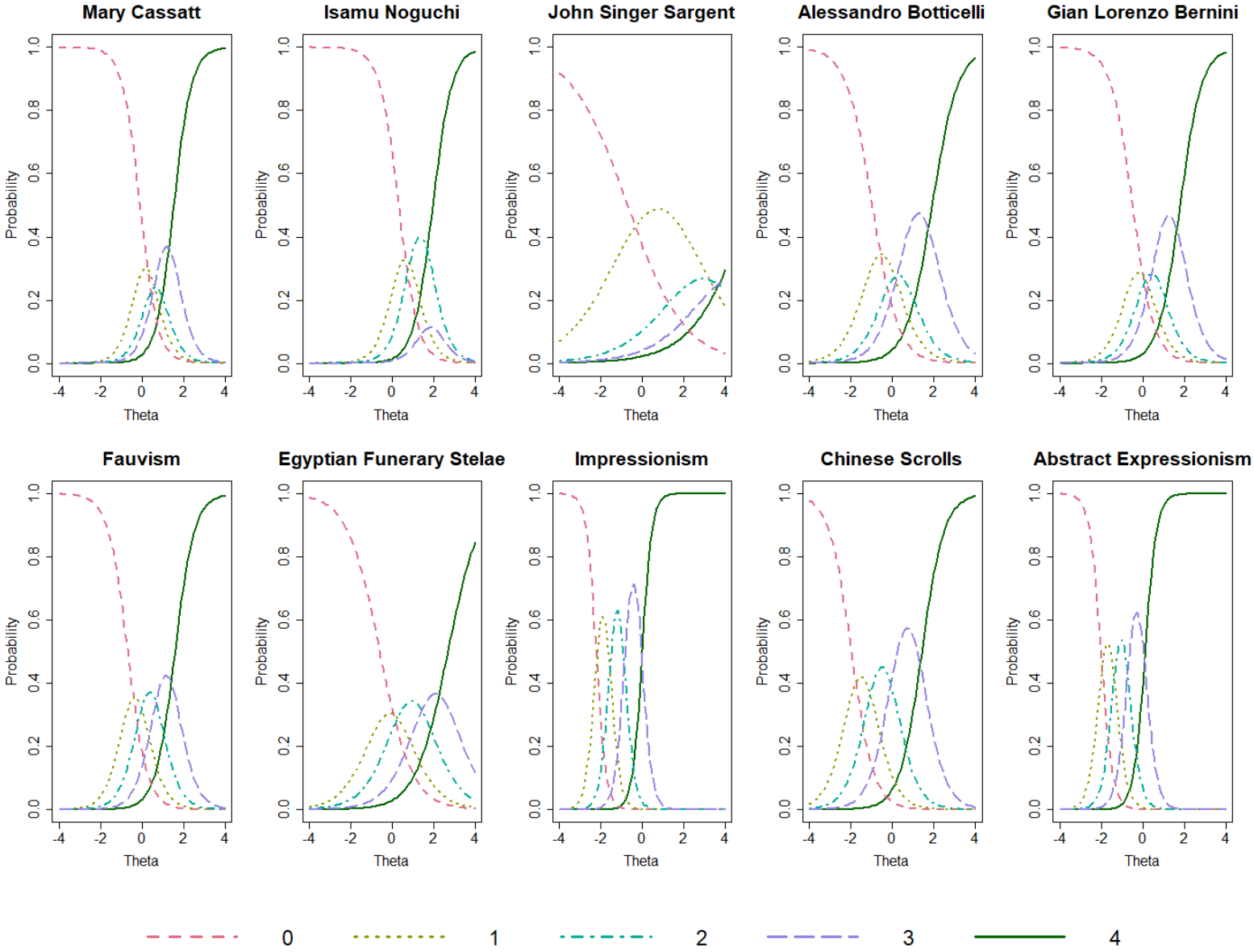

Figure 4. Item information for artistically engaged and non-arts student samples

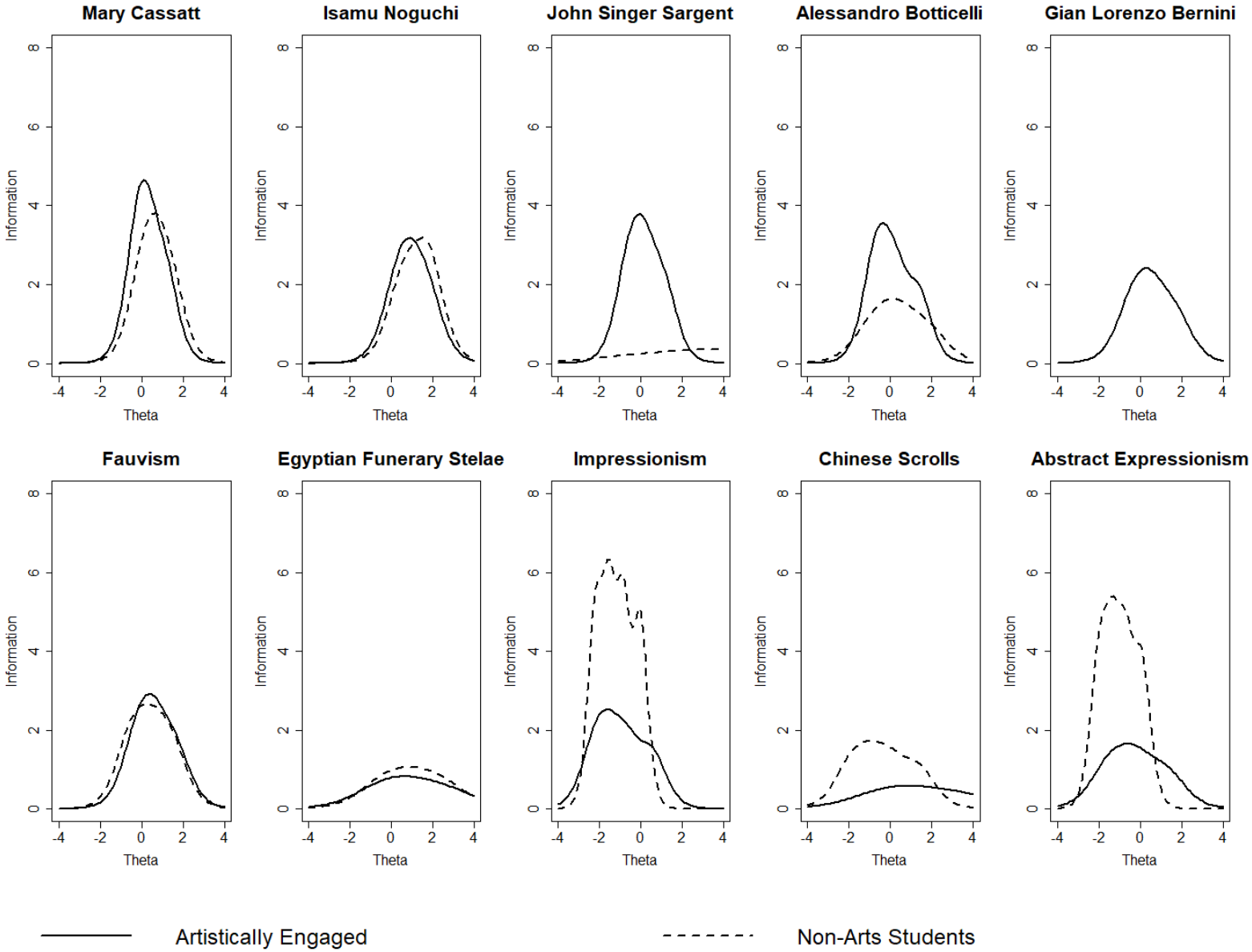

Figure 5. Test information functions