Personality and Individual Differences 117 (2017) 91–100

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Measuring "Good Taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R)

Nils Myszkowski [a,*], Martin Storme [b]

[a] Department of Psychology, Pace University, United States
[b] Laboratoire Adaptations Travail-Individu, Université Paris Descartes - Sorbonne Paris Cité, France

ARTICLE INFO

Article history:
Received 31 October 2016
Received in revised form 21 May 2017
Accepted 22 May 2017
Available online 27 May 2017

Keywords:
Aesthetic sensitivity
Taste
Aesthetic ability
Item-response theory
EFA
Factor analysis

ABSTRACT

Since Eysenck's (1940) discovery of the general factor of visual aesthetic judgments – which he coined "T" (for good *Taste*) – attempts to build a robust measure for it have been largely unsuccessful. The Visual Aesthetic Sensitivity Test (Götz, 1985) is the only "T" measure to have shown acceptable properties, but its structural validity has never been investigated. We randomly split an original sample of 547 adults into two independent samples, to 1) explore the factor structure and revise the VAST, and to 2) cross-validate the structural validity of the revised VAST. Based on EFA and IRT modeling, we found in the first sample that the VAST had unacceptable unidimensionality (McDonald's $\omega_h = 0.59$) and structural validity (CFI = 0.84, RMSEA = 0.04, SRMR = 0.07 for the IRT-3PL fit). After revising the instrument from the factor loadings observed in Sample 1, we tested in Sample 2 the structural validity and unidimensionality of the revised VAST (VAST-R), which was found to have a substantially improved unidimensionality (McDonald's $\omega_h = 0.86$) and structural validity (CFI = 0.95, RMSEA = 0.03, SRMR = 0.07). Further recommendations about the use of the VAST-R are discussed.

1. Introduction

A lot of the research on aesthetic judgments has been focusing on how to empirically define beauty (e.g. Enquist & Arak, 1994; McManus, 2005; Silvia & Barona, 2009), attempting to identify consensus in our judgments in favor of specific features, such as symmetry or complexity. From that experimental stimulus-centered perspective, differences between individuals' judgments are errors that slow research in its discoveries of stable general patterns. Oppositely, the *person* perspective (Jacobsen, 2006) is directly interested in understanding individual differences in aesthetic appreciation. In other words, it represents the differential psychology approach to aesthetic judgment. Aesthetic ability, the extent to which individuals are able to form appropriate or expert aesthetic judgments, is one of its centers of focus that has recently regained interest (e.g., Bezruczko & Frois, 2011; Chamorro-Premuzic & Furnham, 2004; Chatterjee, Widick, Sternschein, Smith, & Bromberger, 2010; Furnham & Chamorro-Premuzic, 2004; Kozbelt & Seeley, 2007; Kozbelt, Seidel, ElBassiouny, Mark, & Owen, 2010; Myszkowski, Çelik, & Storme, 2016; Myszkowski, Storme, Zenasni, & Lubart, 2014; Myszkowski & Zenasni, 2016; Nodine, Locher, & Krupinski, 1993; Silvia, 2007; Silvia & Nusbaum, 2011; Smith & Smith, 2006; Summerfeldt, Gilbert, & Reynolds, 2015).

Indeed, although researchers may explore the characteristics of individuals who have consensual/expert preferences, they may also focus the ability of recognizing what the consensual/expert judgments are – in other words, investigating the capacity to perform expert/consensus-like aesthetic judgments. This ability – which, in everyday language, we typically call "good taste" (Eysenck, 1983) – has been likened to a manifestation of intelligence in the aesthetic domain (Eysenck, 1940; Götz, 1987; Myszkowski, Çelik, et al., 2016). However, because of the multifacetedness, multi-modality, multi-methodology of art, as well as the width and distinctness of art historical movements and themes, research on the measure of aesthetic ability is lagging far behind that of other aspects of mental ability.

The Visual Aesthetic Sensitivity Test (VAST; Götz, 1985) is one of the rare instruments designed for this ability to perform expert-like aesthetic judgments, and the only one with investigated psychometric properties. Yet, empirical investigations about its structural validity are very limited: the present study aims at addressing this issue.

1.1. Why study "good taste"?

With the exception of the musical domain, where "absolute pitch" has been heavily studied (Bachem, 1937; Levitin & Rogers, 2005), researchers rarely study differences in aesthetic judgments under an "ability" framework. In other terms, they focus mainly on what we

* Corresponding author.
E-mail address: nmyszkowski@pace.edu (N. Myszkowski).

http://dx.doi.org/10.1016/j.paid.2017.05.041
0191-8869/

may refer to as "personal taste", but rarely on "good taste" – originally coined the "T" factor (Eysenck, 1940, 1983).

While it is not the case in philosophy and sociology, the literature on good taste in psychology is indeed quite rare, and, apart from a rather recently renewed interest (Bezruczko & Frois, 2011; Chamorro-Premuzic & Furnham, 2004; Furnham & Chamorro-Premuzic, 2004; Myszkowski & Zenasni, 2016; Myszkowski et al., 2014), quite old. The psychometric approach to "T" is even less studied, and the main measures of "T" in the visual domain are at least 30 years old (Götz, 1985; Graves, 1948; Meier, 1940, 1963) – and it is not for the reason that the existing ones had unsurpassable psychometric qualities (Bezruczko & Frois, 2011; Gear, 1986). One can imagine multiple reasons for this past disregard, the main one being that studying good taste seems to come with a bitter – albeit philosophical – pill to swallow: in order to study good taste, such a construct should be meaningful, thus aesthetic standards should exist. Yet, we can make the point that, like intelligence, good taste is for psychologists a scientific object that needs not *exist* per se, but that should be *useful* to help explain phenomena (Myszkowski & Zenasni, 2016). In this regard, the empirical discovery of "T" as a psychological construct – which resembles the early discovery of *g* – explains how "good taste" can be a useful construct.

1.1.1. "T" as an empirical discovery

Like *g*, the general factor of intelligence, "T" has primarily empirical foundations, notably through the early works of Eysenck (Myszkowski, Storme, & Zenasni, 2016). His original study (Eysenck, 1940) on the topic consisted in applying factor analysis to visual aesthetic preferences. More specifically, he asked participants to rank visual aesthetic stimuli, in sets of various categories (portraits, photographs of statues, landscape paintings, photographs of modern steamships, reproductions of pottery, etc.). A first level of factor analysis revealed that a consensus could be observed in favor or disfavor of certain stimuli in the different categories; a second level of factor analysis revealed that the individuals agreed or disagreed with the consensual rankings *consistently* across categories. This finding was interpreted as a general "good taste" – "T" – factor (Eysenck, 1983), which would distinguish individuals depending on the extent to which they agreed with standard judgments of aesthetic stimuli. Additionally, he identified a factor that would explain why some individuals would favor stimuli that are more complex, and he called that disposition the "K" factor

The discovery of both "T" and "K" was very influential. The "K" factor discovery has predominantly triggered interest among creativity researchers. Notably, "K" has been investigated as a feature of creative thinking and artistry (Bezruczko, Manderscheid, & Schroeder, 2016; Bezruczko & Schroeder, 1994; Eysenck & Furnham, 1993; Rosen, 1955). Oppositely, the discovery of "T" has initiated investigations of individual differences in the ability to judge and cognitively process aesthetic stimuli. In relation with its variety of approaches, "T" has been referred to with various terms, notably as "aesthetic sensitivity" (Child, 1964; Frois & Eysenck, 1995; Götz, 1985; Myszkowski et al., 2014; Summerfeldt et al., 2015) and "aesthetic judgment" (Chamorro-Premuzic & Furnham, 2004) – although that term is also used for "K" (Bezruczko, 2013; Bezruczko et al., 2016). Although Eysenck's original experiment defined "T" as the tendency to form consensual judgments, the definition of "T" has evolved as the ability to respond to aesthetic stimuli in manner consistent with "external standards" (Child, 1964, p. 49). In research, such standards have been operationalized both theoretically – through objective aesthetic properties, like balance, symmetry or complexity (Graves, 1948, 1951; Wilson & Chatterjee, 2005) – and empirically – through consensual and/or expert judgments (Götz, 1985; Götz, Borisy, Lynn, & Eysenck, 1979).

1.1.2. The place of "T" in psychological research

As pointed earlier, similar to the *g* factor (general factor of intelligence), the existence of the "T" factor is more of a philosophical question than a psychological one (Myszkowski & Zenasni, 2016), but it is arguable that it is a useful construct of study for a variety of reasons, and a variety of psychological domains, notably aesthetics, creativity, personality and intelligence.

1.1.2.1. "T" and empirical aesthetics. First, as we just noted, "T" is originally a latent dimension, that was found to be a factor explaining aesthetic preference judgments. It is thus directly useful as an explanatory factor in of aesthetic preferences. Additionally, "T" was found to be moderately positively related to creative thinking in the visual domain (Myszkowski et al., 2014), which makes it a useful object of study for creativity researchers, as it may be that the abilities used to create and perceive aesthetic objects overlap. Interesting views on this overlap are notably provided in Tinio's (2013) *Mirror Model of Art* – where art reception is conceptualized as the reversed process of art creation – in Kozbelt's investigations of how artists are advantaged in creation by their perceptual skills (Kozbelt & Seeley, 2007; Kozbelt et al., 2010), in Finke, Ward and Smith's Creative Cognition approach (Finke, Ward, & Smith, 1996; Ward, 2007) approach – which notably states that creative thinking relies considerably on wanderings and associations in a preexisting network of concepts acquired through previous experience, and more generally, in research that highlights the evaluation phases involved in the creative process (e.g., Botella et al., 2013; Cropley, 2006; Silvia, 2008).

1.1.2.2. "T" and personality. Although conceptualized as an ability and not a personality trait, "T" has been found to be related to art interests (Summerfeldt et al., 2015), as well as to emotion-related traits, like sensation-seeking, openness to feelings and openness to fantasy (Myszkowski et al., 2014): it can thus be argued that "T" could relate to a form of interest, empathy or emotional arousal for aesthetic objects – what is often referred to as "aesthetic chills" (Silvia & Nusbaum, 2011). Marketing science has especially been interested in likening "T" to a personality trait, notably focusing on how it can predict the urge to buy well-designed products (CVPA; Bloch, Brunel, & Arnold, 2003; Myszkowski & Storme, 2012).

1.1.2.3. "T" and intelligence. Finally, even though previous conclusions seemed to diverge on the question (Bezruczko & Frois, 2011), a recent meta-analysis (Myszkowski, Çelik, et al., 2016) of the 23 studies presenting correlations between intelligence tests and performance in visual "T" tests concluded that "T" measures were correlated at 0.30 with *g*. Although empirical evidence on the common cognitive processes involved in "T" and *g* is lacking, it was advanced that part of the "T"-*g* relation can be explained by attention shifting, reflective processing, goal management and abstraction.

"T" has certainly recently regained a lot of research interest in various fields, but there still remains a considerable psychometric challenge in measuring good taste (Myszkowski & Zenasni, 2016).

1.2. The challenges of building a visual "T" measure

1.2.1. The multifacetedness of "T"

A major challenge in the measure of "T" is that there is no clear evidence, nor conceptual framework, of its dimensionality. "T" has been originally conceptualized as unidimensional (Eysenck, 1940), but this conceptualization has been questioned, for the reason that there are many different ways that "T" is assessed. Indeed, mimicking the *g*-to-IQ shift in intelligence measurement, recent conceptions of aesthetic ability have pointed out the need to consider a multifaceted "Aesthetic Quotient" (AQ) construct – defined as the "global capacity to identify, explore, understand, seek stimulation in and respond to the elements, composition and meaning of art and aesthetic objects" (Myszkowski & Zenasni, 2016, p. 2). It has also been suggested (Myszkowski, Storme, et al., 2016) that "T" and "K" could, similar to fluid-crystallized intelligence theories (Cattell, 1963), be respectively likened to fluid/

performance and crystallized/verbal aesthetic abilities. In any case, whether the narrower "T" or the broader "AQ" approach of aesthetic ability is considered, there is little doubt that "T" has, because of its definition as the ability to form "standard" (expert-like, consensus-like, etc.) judgments of taste (Child, 1964), a central position in aesthetic ability (Myszkowski & Zenasni, 2016). Therefore, the availability of "T" measures that are psychometrically robust is of paramount interest to the scientific study of aesthetic ability, whether a multidimensional or a general factor is favored.

1.2.2. Typical construction process

Recent research on "T" (Bezruczko & Frois, 2011; Chamorro-Premuzic & Furnham, 2004; Furnham & Chamorro-Premuzic, 2004; Myszkowski et al., 2014; Summerfeldt et al., 2015) has particularly used the Meier Art Tests (Meier, 1940, 1963), the Judgment Design Test (Graves, 1948) and the Visual Aesthetic Sensitivity Test (VAST; Götz, 1985) to measure it. While they have different material – figurative artworks for the Meier Art Tests, basic geometrical designs for the Design Judgment Test, and informal abstract art works for the VAST – the three of them include exclusively black and white stimuli, and function using the same item construction paradigm: "controlled alteration" (Meier, 1928, p. 188). It consists in altering a stimulus (typically, an existing artwork) in order to create one or multiple alternate versions of lesser aesthetic quality than the original. The task completed by the test taker consists in identifying which of the versions is the one of the best aesthetic quality. The scoring system consists in counting the number of correct identifications of the test taker.

1.2.3. Psychometric investigations

In the controlled alteration approach, the content validity of the items is supposed to be ensured by the item creation process itself, and, for the Meier Art Tests and the Judgment Design Test, this is the only element of content validity available, in the sense that can only rely on the test authors' expertise. For the VAST, however, and it is one of the reasons to conduct this research using this test rather than the other two, content validity is not only ensured by the controlled alteration process: it is also verified through both the consensual agreement of a community sample and the unanimous agreement of eight "well-known" (Götz et al., 1979, p. 796) art experts.

Additionally, as far as the Meier Art Tests and the Judgment Design Test are concerned, psychometric investigations are largely non-existent, or point out to major faults in terms of structural validity (Eysenck, 1967), content validity (Götz & Götz, 1974) as well as predictive validity – the Design Judgment Test notably failing to differentiate between artists, art students and non-students (Eysenck, 1972; Eysenck & Castle, 1971). In contrast, the VAST has been the object of numerous psychometric investigations, in terms of internal consistency, external criterion validity, and cross-cultural measurement invariance (Chan, Eysenck, & Götz, 1980; Eysenck, Götz, Long, Nias, & Ross, 1984; Frois & Eysenck, 1995; Götz, 1987; Iwawaki, Eysenck, & Götz, 1979; Myszkowski et al., 2014), making it the only psychometrically recommendable visual "T" measure (Myszkowski et al., 2014).

1.2.4. The structure of the VAST

In spite of its empirical content validity investigation, the structural validity of the VAST has never been investigated. Satisfactory Cronbach's αs were previously reported (Myszkowski et al., 2014), but, although α is often reported as an estimate of the homogeneity – or unidimensionality – of an instrument, in reality, Cronbach's α does not express characteristics of an instrument's factor structure (Revelle & Zinbarg, 2009; Sijtsma, 2009; Zinbarg, Yovel, Revelle, & McDonald, 2006). Consequently, reports of the VAST's Cronbach's α cannot provide information about its factor structure. Because the factor structure of the VAST, theoretically unidimensional, has never been empirically investigated, because of the very complex nature of its aesthetic material, and because other "T" measures – notably the Judgment Design Test (Eysenck, 1967) – have been shown to not demonstrate their claimed unidimensionality empirically, we decided to design a study that would allow to investigate the homogeneity of the VAST, and to carve in the VAST a revised version (the VAST-R), with a finally appropriate structural validity.

1.3. The aim of this study

In this study, we aimed at investigating and reinforcing the unidimensional structure of the VAST – thus creating a VAST-R – by using a homogeneous subset of items, rather than the entire test. Because of this two-stage aim, our statistical investigations were conducted in two successive stages: 1) VAST factor structure investigation/exploration (and subsequent VAST-R construction) and 2) VAST-R factor structure confirmation on an independent sample.

To form two independent samples, we decided to split a sample of VAST respondents in two samples equivalent in sample size and in VAST scores (Hastie, Tibshirani, & Friedman, 2011; Kuhn, 2008). The first sample was used to explore the homogeneity of the original VAST, and to build the VAST-R. The second sample was used to confirm the strengthened unidimensional factor structure.

We hypothesized that 1) the original VAST would show a weak unidimensionality in the first sample (with one general factor that would however not encompass much of the variance of the item scores), but that 2) this analysis would help us build a strong unidimensional VAST-R, with both good internal consistency and a strong unidimensional factor structure, as tested through confirmatory analyses in the independent cross-validation sample.

2. General method

2.1. Participants

The original sample was composed of a total of 547 undergraduate students – 227 males and 320 females, aged between 19 and 25 ($M = 20.5$ years, $SD = 0.96$) – from a French business school. They were all 3rd year college students majoring in general business and management, without any further specialization. They responded the VAST on a voluntary basis, without any compensation or benefit for participating. They all responded on computer, without having received any previous introduction to or training in visual art, empirical aesthetics or psychometrics.

2.2. Instruments

The participants all responded the (original) Visual Aesthetic Sensitivity Test (VAST; Götz, 1985). The VAST is a 50-item measure of the "T" factor in the visual domain (Eysenck, 1983). Following the controlled alteration procedure earlier explained in our review of visual "T" measures, each item consists of a pair of black and white paintings, one of which being an altered version of the other one. It is altered so that it is supposed to have lower aesthetic quality. The stimuli were created by the abstract art painter Karl Otto Götz (example items of the VAST are presented in Fig. 1).

For each item, the test taker has to indicate which of the two paintings, which are presented side by side, is the one of better aesthetic value. Different terms for high or low aesthetic value are used in the instructions (Götz, 1985): "superior from the point of view of design", "more harmonious", "better balanced", "better adapted in the way the elements are ordered, and in the way the lines are drawn", "faults or errors which destroy the balance of the picture", "more balanced", "better formulated", "better", "well ordered and circumscribed figure", "better designed". The participants are also explicitly instructed to not respond which of the two pictures they prefer, but instead which has been better designed. Test takers are also indicated in the instructions that there is, for each item, a correct response determined by art experts (painters

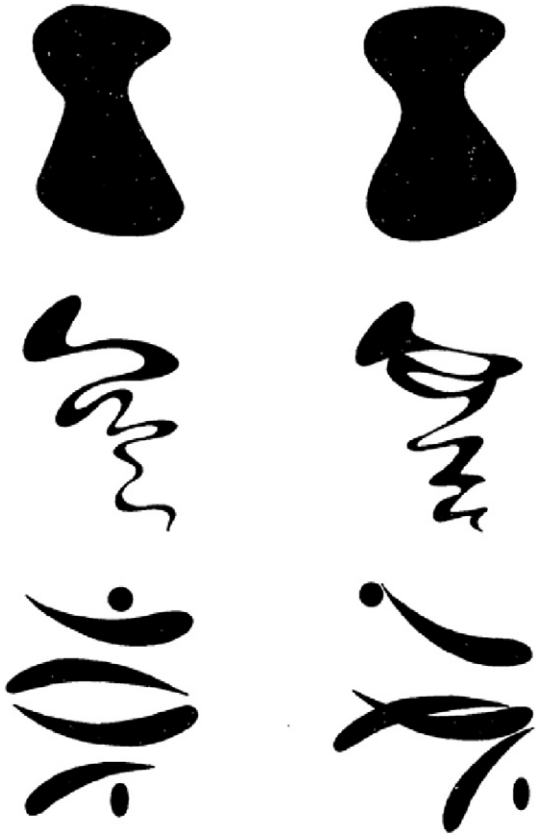

Fig. 1. Example items of the Visual Aesthetic Sensitivity Test (from Götz et al., 1979).

and graphic artists). Finally, test takers are instructed to respond, even when they have difficulties answering.

The count of correct recognitions is used as the VAST score, leading to total scores that theoretically range from 0 to 50. Nevertheless, because of the encouragement to respond even in cases of uncertainty, scores typically range between 25 (the expected score for a participant responding totally randomly) and 50. In this study, we observed a mean score of 35.8 ($SD = 6.1$), which is close to the scores that were previously obtained on undergraduate students (Myszkowski et al., 2014). Although not being a measure of unidimensionality (Dunn, Baguley, & Brunsden, 2014; Revelle & Zinbarg, 2009; Sijtsma, 2009; Zinbarg et al., 2006), a satisfactory Cronbach's α (0.88) was observed (based on the matrix of tetrachoric correlations), which is higher than what was reported in previous studies (Myszkowski et al., 2014).

2.3. Overall data analysis strategy

2.3.1. Data splitting

As explained earlier, our approach consisted of two main phases: 1) Investigating the structural validity of the VAST, and consequently creating the VAST-R, and 2) confirming the good structural validity of the VAST-R. Because here the confirmatory analyses performed in the second step are directly derived from the item selection subsequent to the first exploratory analysis, the same data could not be used. Using the same sample for both steps in our case would present a danger of *overfitting*, meaning that we would end up with a new instrument that would by design have good properties in the sample, but whose properties could not necessarily be generalizable to another sample (Hastie et al., 2011).

For this reason, we first split the original sample, using a *2-fold* data split (also called *holdout*) method for cross-validation (Hastie et al., 2011; Zhang, 1993). To do so, we used the R package 'caret' (Kuhn, 2008; Kuhn & Johnson, 2013). Using this package (heavily used in predictive modeling in R), we partitioned the data into two "equivalent" parts. By equivalent, we mean that observations of the original dataset were classified in one group or the other randomly, but ensuring that the two groups have approximately the same distribution of VAST total scores – thus ensuring groups of similar "T". Concretely, this is achieved by using quotas per quintile. It consists in randomly assigning observations to the groups within each quintile of the distribution of the VAST scores (Kuhn, 2008). As a result, we assumed that the different levels of "T" were equivalently represented in each group. "Sample 1" was used as the subsample used in the first study (exploration of the VAST and construction of the VAST-R), and "Sample 2" was the subsample used in the second study (cross-validation of the structural validity of the VAST-R).

2.3.2. Verifying equivalence between the samples

The assignment process being the result of a randomized assignment from each quintile, the two resulting group sizes were not assumed to be perfectly equal – but at least very similar (Kuhn, 2008). In our case, group sizes were very close ($n_1 = 275$, $n_2 = 272$). The two groups also logically had very close VAST means ($M_1 = 35.8$ years, $M_2 = 35.8$ years, Cohen's $d = 0.01$, $t(545) = 0.11$, $p = 0.91$). Additionally, there were no significant group differences in terms of mean ages ($M_1 = 20.7$ years, $M_2 = 20.8$ years, Cohen's $d = 0.08$, $t(545) = -0.88$, $p = 0.38$) or gender distributions ($n_{1,\ Male} = 120$, $n_{1,\ Female} = 155$, $n_{2,\ Male} = 107$, $n_{2,\ Female} = 165$, Cramer's $V = 0.04$, $\chi^2(1) = 0.87$, $p = 0.35$). This indicates that we had two samples that were similar by different aspects, though independent.

2.3.3. Internal consistency and unidimensionality

In spite of Cronbach's (1951) α being the most largely used measure of "internal consistency" – although the term needs clarification (Sijtsma, 2009) – it does not demonstrate (but assumes) unidimensionality (Revelle & Zinbarg, 2009; Sijtsma, 2009; Zinbarg et al., 2006). Indeed, it has been shown that it is possible to vary a factor structure (from one to multiple factors) while maintaining the same α (Sijtsma, 2009). For that reason, recently, the use of a new measure of unidimensionality has been encouraged: based on Exploratory Factor Analysis general factor loadings and relying on more realistic assumptions (Dunn et al., 2014), McDonald's ω_h (Zinbarg et al., 2006) is "the proportion of variance in the scale scores accounted for by a general factor" (Zinbarg et al., 2006, p. 122), and is therefore a clearly interpretable measure of unidimensionality.

Through our analyses, we reported Cronbach's α and McDonald's ω_h. Both were estimated using the R package 'psych' (Revelle, 2016). From the VAST to the VAST-R, test length was shortened while unidimensionality was (hypothetically) strengthened. Therefore, we did not make hypotheses on a change in α. However, because ω_h is a direct measure of unidimensionality, and because we aimed at improving unidimensionality, we hypothesized an increase in ω_h between the VAST and the VAST-R.

2.3.4. Factor analyses/structural validity

It was clearer to present the factor analyses used in the related section of the study. Nonetheless, more generally, our analysis plan consisted in using Exploratory Factor Analyses (based on the tetrachoric correlation matrix) of the VAST items in the first sample, for the reason that there is no available evidence proving or disproving any structure for this instrument, as we previously explained. IRT analyses followed to confirm the weakness of the unidimensionality of the VAST, thus demonstrating its low structural validity. Then, Exploratory Factor Analysis factor loadings were used to form the VAST-R. Finally, in Sample 2, because we investigated a revised test with a hypothetical strong and known factor structure, we used Confirmatory Factor Analysis, using unidimensional IRT models.

3. Study 1: investigating and revising the VAST

In this study, we first investigated the factor structure of the original VAST (Götz, 1985) on Sample 1. Although theoretically unidimensional, the structure of the VAST hasn't been investigated before, and therefore, there is really no evidence for the unidimensionality of the VAST, which could well have a more complex empirical structure.

A preliminary step of this investigation of structural validity was the examination of internal consistency and unidimensionality through respectively Cronbach's α and McDonald's ω_h. The VAST being quite long (50 items), it logically showed satisfactory internal consistency ($\alpha = 0.89$). However, interestingly, as hypothesized, it showed a weak unidimensionality, as the proportion of variance in the item scores accounted for by a general factor only amounted to 59% ($\omega_h = 0.59$).

The 3 steps of this first study used Sample 1 to perform a more complete evaluation and improvement of the structural validity of the VAST, by exploring the structure of the VAST through EFA (Step 1), evaluating the fit of confirmatory unidimensional models to the data (Step 2), and building a revision of the VAST (Step 3).

3.1. Step 1: exploring the factor structure of the VAST

3.1.1. Method

Based on the matrix of tetrachoric correlations between all the 50 items of the VAST, we first conducted an Exploratory Factor Analysis (EFA) with parallel analysis (Hayton, Allen, & Scarpello, 2004; Horn, 1965), using the R package 'psych' (Revelle, 2016).

3.1.2. Results

As hypothesized, the factor structure of the VAST did not appear unidimensional: the parallel analysis indicated to retain 15 factors, and 8 factors had eigenvalues above 1. However, the scree plot inspection (Cattell, 1966) revealed a large drop in eigenvalues starting at the second factor (from an eigenvalue of 8.56 for the first factor, to eigenvalues below 2.68 for the remaining retained factors). The scree plot with the observed eigenvalues, along with the resampled eigenvalues from the Parallel Analysis, is presented in Fig. 2.

Additionally, the examination of the loadings of the items on these remaining factors did not indicate interpretable factors, meaning that they did not appear to relate to any specific item content. From our EFA analysis, we concluded that there was only one meaningful factor to extract, but that, because of the number of factors that the parallel analysis suggested to retain, the test could present a weak unidimensionality, and therefore bad structural validity, which was in line with our first hypothesis.

3.2. Step 2: confirming the weakness of the VAST's unidimensionality

To know if that apparent low unidimensionality indicated poor constructed validity, we evaluated the fit of unidimensional factor structures – using IRT modeling – to the VAST item scores.

3.2.1. Method

Item-Response Theory (IRT) is a psychometrical framework for modeling the relationship between a test taker's underlying latent trait level (in our case, what we assume to be "T") and his/her item responses (Chalmers, 2012). It more specifically models the probability – for dichotomous items like the VAST items, typically, through logistic models – of responding in a specific way – typically, the probability to answer correctly – for different latent trait levels (named θ).

IRT logistic models differ according to the number of parameters they use, and that number of parameters is typically set to account for the test's characteristics (for example, the possibility for test takers to guess the correct responses, or the probability that test takers are

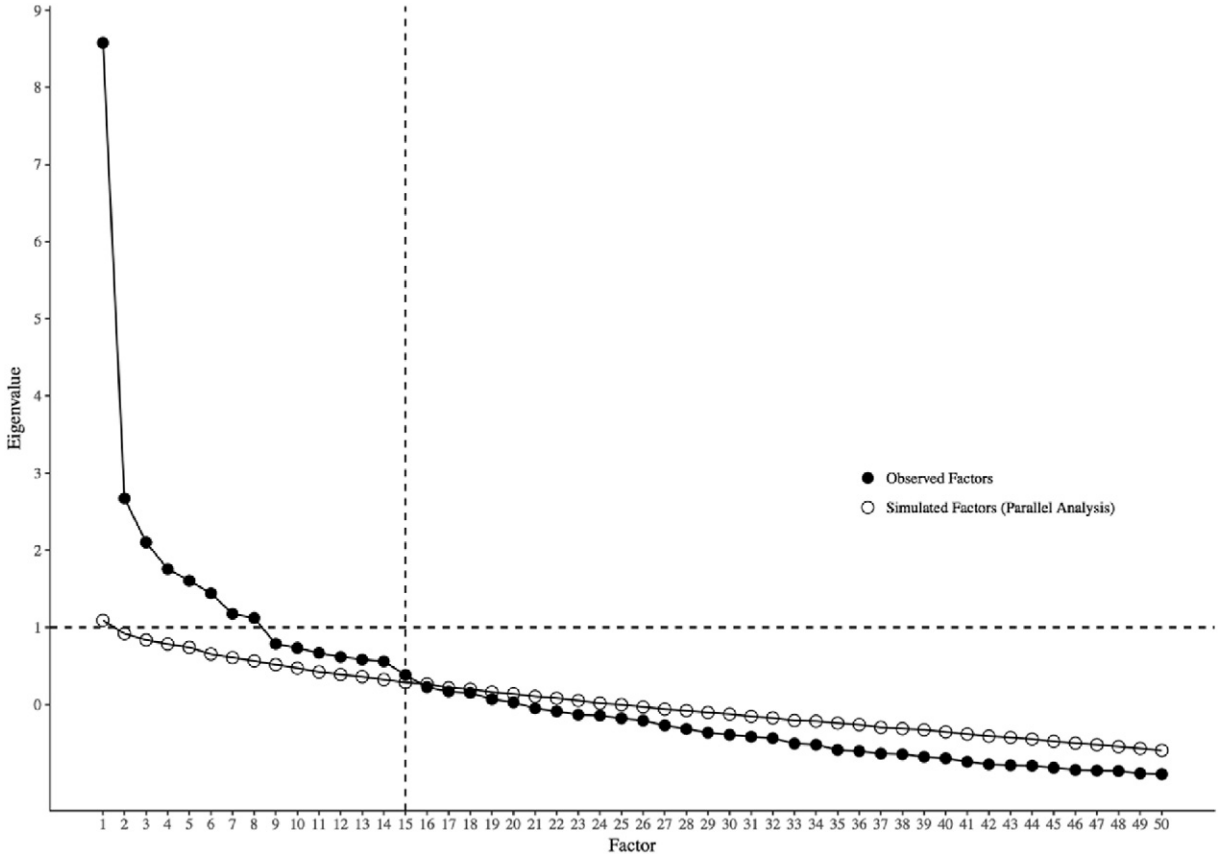

Fig. 2. Scree plot of the Exploratory Factor Analysis (with Parallel Analysis).

distracted). A one-parameter logistic model (1PL) assumes that items measures equally well the trait, and only vary in difficulty (the value of θ for which the slope of the logistic function is maximized); a two-parameter model (2PL) allows for items to also vary in discrimination (the slope of the logistic function); a three-parameter logistic model (3PL) adds a lower asymptote to the logistic function (Birnbaum, 1968), accounting for varying degrees of pseudo-guessing; a four-parameter logistic model (4PL) adds an upper asymptote to the logistic function (Barton & Lord, 1981), to account for varying degrees of "distraction" or "carelessness".

We decided not to eliminate the possibility that test takers try to guess the correct response, or that individuals could make inattention mistakes. For this reason, we tested a variety of IRT dichotomous models: 1PL, 2PL, 3PL and 4PL. As a weak unidimensionality was hypothesized, we anticipated that all of these models would only provide a mediocre fit of the data.

The IRT models were fit using the R package 'mirt' (Chalmers, 2012). As recommended (Hu & Bentler, 1999), to evaluate the fit of models to the data, we used the Standardized Root Mean Square Residual (SRMR) with a cut-off of 0.08, the Root Mean Square Error of Approximation (RMSEA) with a cut-off of 0.06, and the Comparative Fit Index (CFI) with a cut-off of 0.95.

3.2.2. Results

As hypothesized, none of the IRT models had a satisfactory fit to the data (the fit indices are reported in Table 1). This indicates that, as hypothesized, the structure of the VAST is only weakly unidimensional, thus allowing concluding that the VAST has unsatisfactory structural validity.

3.3. Step 3: creation of the VAST-R

Since the VAST did not show adequate structural validity, we built its revision, the VAST-R, with the objective of strengthening its structure. IRT parameter estimates would have constituted an appropriate basis for item selection. However, the instructions of the VAST overtly encourage pseudo-guessing, which is only modeled in 3PL and 4PL models, and, with 50 items, the low sample-size-to-number-of-parameters ratio of the 3PL and 4PL models led us to consider the possibility that many IRT parameters estimates could in this case be too unstable. We thus made the decision to instead use EFA loadings to select the VAST-R items. In this view, we ran another EFA on Sample 1, but this time extracting only one factor. Our aim here was not to investigate the structure anymore, but to select items using factor loadings.

The unidimensional IRT models fitting very poorly the data, we used the EFA factor loadings, which offer clear criteria and rules of thumb for keeping or discarding items (Kline, 2005). In line with typical recommendations (Floyd & Widaman, 1995), items with factor loadings above 0.4 were kept in the VAST. These amounted to 25 items, which represents a total of half the original VAST.

Table 1
IRT fit indices of the VAST (Sample 1 and full original sample).

Model	Sample	χ^2	*df*	CFI	SRMR	RMSEA	AICc
1PL	Sample 1	2486.12	1224	0.416	0.100	0.061	14,699.3
	Full original sample	3235.23	1124	0.510	0.085	0.055	29,253.1
2PL	Sample 1	1958.42	1175	0.781	0.074	0.049	14,612.2
	Full original sample	2641.09	1175	0.766	0.061	0.048	28,938.8
3PL	Sample 1	1683.38	1125	0.844	0.073	0.043	14,867.7
	Full original sample	2320.64	1125	0.809	0.061	0.044	29,007.6
4PL	Sample 1	1710.67	1075	0.822	0.076	0.046	15,548.32
	Full original sample	2179.50	1075	0.824	0.061	0.043	29,010.5

Note. CFI – Comparative Fit Index; SRMR – Standardized Root Mean Square Residual; RMSEA – Root Mean Square Error of Approximation; AICc – Akaike Information Criterion (corrected).

3.4. Discussion

This first study allowed us to confirm that the original VAST had a weak unidimensional structure, and thus a low structural validity. Through EFA factor loadings, we were able to carve in the VAST a revised version, the VAST-R, which we anticipated would have a stronger unidimensional structure, and thus a more robust structural validity.

4. Study 2: evaluating the structural validity of the VAST-R

In this second study, we used an independent sample (Sample 2), to investigate the psychometric properties of the newly constructed VAST-R.

Here again, a preliminary step of this investigation of structural validity was the examination of internal consistency and unidimensionality through respectively Cronbach's α and McDonald's ω_h. Albeit half the length of the VAST, the VAST-R still had satisfactory internal consistency in this cross-validation sample, with a Cronbach's α of 0.87, which is very close to that of the original VAST, previously reported. However, a substantial increase in unidimensionality was found in the VAST-R, with a ω_h of 0.86, which means that the general factor explained much more (86% for the VAST-R vs. 59% for the VAST) of the variance of the item scores.

The main aim of this study was to confirm the expected good structural validity of the revised VAST (VAST-R), through confirmatory factor analyses using IRT Modeling.

4.1. Method

To confirm the unidimensional factor structure, IRT-1PL, 2PL, 3PL and 4PL models were fit to the data using the R package 'mirt' (Chalmers, 2012). Like before, to evaluate the fit of models to the data, we used the Standardized Root Mean Square Residual (SRMR) with a cut-off of 0.08, the Root Mean Square Error of Approximation (RMSEA) with a cut-off of 0.06, and the Comparative Fit Index (CFI) with a cut-off of 0.95. Likelihood Ratio tests and corrected Akaike Information Criteria (AICc) were used to compare the fit of the different models (a lower value indicates a more parsimonious fit).

4.2. Results

As hypothesized, the fit indices of the 3PL and 4PL models were satisfactory. Indicating that the VAST-R had good structural validity and a strong unidimensionality – Table 2 presents a complete report of the various fit indices in Sample 2, as well as in the entire original sample.

In the entire sample, the Likelihood Ratio tests favored the 4PL model, which fit significantly better than all other models (all $p < 0.001$). More specifically, the 2PL model fit the data significantly better than the 1PL model ($\chi^2(24) = 146.59$, $p < 0.001$); the 3PL model fit the data marginally significantly better than the 2PL model ($\chi^2(25) = 35.21$, $p = 0.08$) and significantly better than the 1PL model ($\chi^2(49) = 181.80$, $p < 0.001$); the 4PL model fit the data significantly better than the 3PL

Table 2
IRT fit indices of the VAST-R (Sample 2 and full original sample).

Model	Sample	χ^2	*df*	CFI	SRMR	RMSEA	AICc
1PL	Sample 2	430.04	299	0.843	0.100	0.040	6316.1
	Full original sample	557.60	299	0.865	0.088	0.040	12,395.6
2PL	Sample 2	373.90	275	0.929	0.066	0.036	6269.3
	Full original sample	503.69	275	0.930	0.055	0.039	12,304.6
3PL	Sample 2	325.38	250	0.946	0.065	0.033	6329.8
	Full original sample	420.54	250	0.948	0.056	0.035	12,333.3
4PL	Sample 2	316.62	225	0.934	0.065	0.039	6386.8
	Full original sample	380.18	225	0.953	0.055	0.036	12,306.2

Note. CFI – Comparative Fit Index; SRMR – Standardized Root Mean Square Residual; RMSEA – Root Mean Square Error of Approximation; AICc – Akaike Information Criterion (corrected).

model ($\chi^2(25) = 98.08$, $p < 0.001$), the 2PL model ($\chi^2(50) = 133.29$, $p < 0.001$) and the 1PL ($\chi^2(74) = 279.9$, $p < 0.001$).

4.3. Discussion

As opposed to the results of Study 1, which demonstrated the poor structural validity of the VAST, these results demonstrate the good structural validity of the VAST-R, advocating for its use. They also show that a good internal consistency was maintained in this revision.

5. Additional analyses

To give further recommendations about the use of the VAST-R, we performed additional analyses. These analyses being separate from our hypotheses, they were performed, for improved accuracy, on the entire sample.

5.1. Asymptotes

IRT-3PL and 4PL models fitted the data better than 1PL and 2PL models, indicating that at least a pseudo-guessing asymptote should be taken into account in investigations of the VAST-R. However, the item response functions of the IRT-3PL model – presented in Fig. 3 – notably show that, regarding pseudo-guessing, interestingly, items for the most part had either a pseudo-guessing estimate of about 0.50 (which is in line with a purely random guess) or 0 (indicating that participants for these items "actively" pick the "wrong" answer).

Such bimodality of the pseudo-guessing parameters may indicate two different processes for responding, and therefore potentially two types of items. More specifically, we may suggest that, for the items with a pseudo-guessing of 0.5, individuals succeed because they detected the elements of superior quality on the better-designed version. In that case, individuals who fail an item fail because they did not detect that element, and as a consequence had to guess at random (as a reminder, they are encouraged to respond even when not sure). Oppositely, for the items with a pseudo-guessing of 0, we may advance that individuals are "trapped" by specific details of the stimuli that they mistake for an indication of the correct (or incorrect) response. In other terms, it is possible that individuals who succeed use pertinent signals, while individuals who fail use non-pertinent signals.

Regarding upper asymptotes, even though the fit of the IRT-3PL model appeared better than all the other models in the cross-validation sample, we can note that 1) in the cross-validation sample, the IRT-4PL model – which adds to the IRT-3PL a parameter of upper asymptote – still had a borderline acceptable fit, and that 2) in the entire sample, the same IRT-4PL actually also had a satisfactory fit to the data – it even had a slightly better fit than the IRT-3PL model.

To make more robust recommendations regarding which IRT model to use with the VAST-R (both for scoring and for further structure investigations), we investigated the different IRT models further, using two

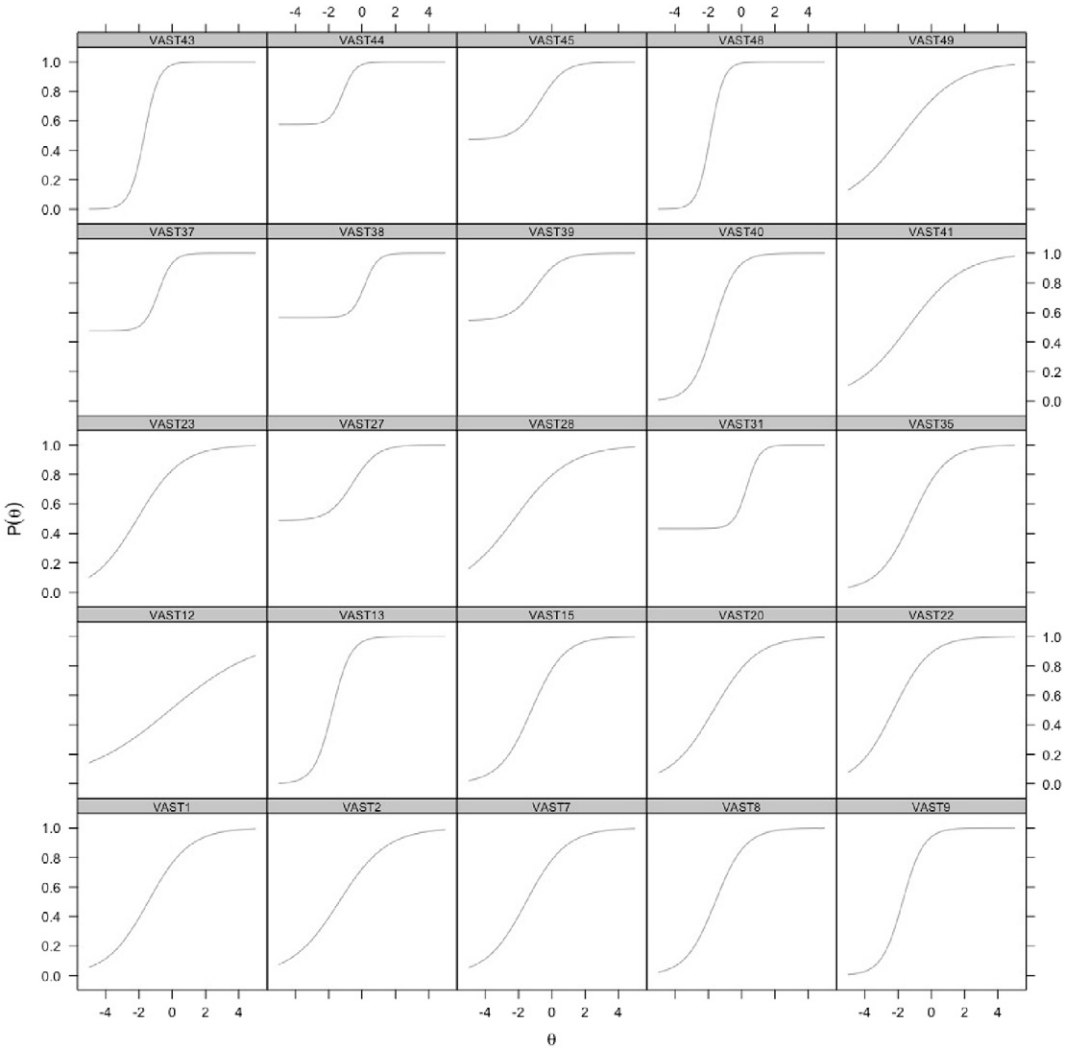

Fig. 3. Item Response Functions of the items of the VAST-R.

cross validation methods: *k-fold* cross-validation and bootstrap resampling. For *k-fold* cross-validation, we first randomly split the original sample into 10 independent samples, using the procedure described earlier. We then trained 4 IRT models (1PL, 2PL, 3PL and 4PL) on a first sample. Afterwards, we fixed the parameters estimated in the training sample, and fit the 4 IRT models on each of the remaining 9 samples. For each sample and each model, we computed the reliability of the scores through both empirical reliability – which is the reliability in the sample distribution of the ability estimates – and marginal reliability – which is the reliability in an assumed prior distribution of ability estimates (here, we used a normal prior distribution). These reliability estimates were then averaged across the 9 testing samples. Overall, as expected from the results of Study 2, the model that provided the most reliable ability estimates was the 3PL model, when looking at both empirical reliability ($r_{xx,1PL} = 0.68$, $r_{xx,2PL} = 0.72$, $r_{xx,3PL} = 0.82$, $r_{xx,4PL} = 0.82$) and marginal reliability ($r_{xx,1PL} = 0.73$, $r_{xx,2PL} = 0.70$, $r_{xx,3PL} = 0.79$, $r_{xx,4PL} = 0.73$). For bootstrap resampling, we estimated the IRT models on the full original sample. We then fixed the estimated parameters and fit the IRT models on 1000 bootstrap resamples. We then averaged the reliability estimates obtained over the 1000 resamples. Consistent with the results of the *k-fold* validation, the 3PL provided the highest average empirical ($r_{xx,1PL} = 0.72$, $r_{xx,2PL} = 0.72$, $r_{xx,3PL} = 0.78$, $r_{xx,4PL} = 0.76$) and marginal reliability ($r_{xx,1PL} = 0.74$, $r_{xx,2PL} = 0.71$, $r_{xx,3PL} = 0.77$, $r_{xx,4PL} = 0.77$).

5.2. Using sum scores

Deriving latent scores from IRT modeling certainly presents advantages in terms of accuracy and conditional reliability, but it also presents disadvantages in terms of practicality, in that it is demanding in terms of sample size (parameter estimations may notably fail to converge with small sample sizes), and capabilities of the statistical package used. For that reason, we extracted the θ scores from the IRT-3PL model, and computed the correlation coefficient between them and the scores computed by simply summing the 25 item scores. This correlation was nearly perfect ($r = 0.96$, $t(545) = 81.0$, $p < 0.001$), indicating that sum scores can be nearly perfect substitutions for IRT-3PL factor scores.

6. General discussion

Our hypothesis about the original test was confirmed in Study 1: the VAST seemed indeed somewhat unidimensional, in that, contrary to the Design Judgment Test (Eysenck, 1967), only one general factor was interpretable from the EFA. Nevertheless, as hypothesized, its unidimensionality was weak, as one-dimensional IRT-1PL, 2PL, 3PL and 4PL models all demonstrated unacceptable fit, and as McDonald's unidimensionality estimate ω_h was mediocre. From the loadings of a unidimensional EFA, we selected a subset of VAST items to be included in the VAST-R, in the hope that this revised version would show better structural validity.

In Study 2, on an independent sample, as hypothesized, the unidimensionality of the VAST-R was found to be much stronger than what was observed for the VAST in Study 1, with notably a unidimensional IRT-3PL confirmatory model fitting satisfactorily the data (Hu & Bentler, 1999), and a substantially increased McDonald's unidimensionality estimate ω_h of 0.86, indicating an improved and now satisfactory structural validity for the VAST-R. Internal consistency (evaluated through Cronbach's α) was kept at about the same level as that of the VAST, for a measure of half its length.

In supplementary analyses on the entire sample, we notably showed a dichotomy between items that have a pseudo-guessing estimate of 0, and items that have a pseudo-guessing estimate of 0.5. As we explained, this dichotomy may be a sign that there are actually two different cognitive processes that underlie the performance at these two types of items. Although it may seem like a psychometric detail, it may be the tip of the iceberg, as it may reveal two different "good taste" mechanisms, levels of processing, or even definitions. Indeed, in one case, it seems here that high "T" corresponds to an ability to detect a signal – which is in line with a more perceptual explanation of art expertise (Kozbelt & Seeley, 2007; Kozbelt et al., 2010) – whereas in the other, "T" is about correctly judging of the pertinence of a signal that has been detected – which corresponds more to an inhibition or attention-shifting mechanism, a recently proposed "T" related cognitive process (Myszkowski, Çelik, et al., 2016). We may here indeed, in spite of a unidimensional measurement instrument, be evaluating an aptitude that taps into distinct cognitive processes. Although we did not, from the inspection of the related items, detect any stylistic particularity that would distinguish these two types of items, we would recommend that future research investigates this, for example by debriefing with the participants, or using eye-tracking to understand which cue they based their decision on. On one hand, for the first type of item – detecting vs. not detecting a signal – we would expect failing participants to not focus on any specific signal, thus finally responding randomly, while succeeding participants would have focused on that signal. On the other hand, for the second type of item – detecting a pertinent vs. non-pertinent cue – we would expect that participants who failed focused on a misleading detail, whereas the participants who succeeded accurately judged the detail as not pertinent.

Even though further cross-validation techniques (*k-fold* and bootstrap) encouraged the use of a 3PL model, the comparison of fit between IRT models in the entire sample revealed that the 4PL model could actually fit VAST-R scores better than the 3PL model. As a consequence, we suggest that no conclusion is drawn about the comparison of the fit between 3PL and 4PL modeling on the VAST-R. Indeed, both seemed here to fit almost equivalently well, and it seems like, even though the estimation of a variable lower asymptote (which is both in IRT-3PL and 4PL models) seems necessary (from both the VAST-R instructions and the observed fit indices), the estimation of an upper asymptote should still be questioned. Practically speaking, we recommend that, in the absence of further evidence, future IRT investigations consider both IRT-3PL and 4PL modeling, and use the one with a better empirical fit.

We investigated the use of sum scores as substitutions for the more computationally intensive IRT-3PL models. We found a nearly perfect correlation between factor scores from the IRT-3PL model and sum scores, and, consequently, we advise that future studies that do not focus on the structural validity of the VAST-R or that do not necessitate the advantages of IRT modeling (notably conditional standard errors of measurement) use sum scores. In the case of low sample sizes – which tend to complicate IRT parameter estimation processes – sum scores may actually be more reliable, as they will not rely on potentially unstable estimates.

Although we found the VAST-R to have a much more robust structural validity than the VAST, other properties of the VAST-R should now be investigated. Notably, test-retest reliability should be investigated – it has never been investigated in the original VAST or any "T" measure – as well as criterion validity. More specifically, future research should verify that the VAST-R is related – as found for different "T" measures in a recent meta-analysis (Myszkowski, Çelik, et al., 2016) – to measures of intelligence. In fact, future research could investigate whether the VAST-R, having now improved psychometric properties over previous "T" measures, has also more substantial – because less attenuated – relations with intelligence measures that were previously found with other visual "T" measures. Such prospective results could actually be extended to the relations between the VAST-R and other related constructs, like personality (Chamorro-Premuzic & Furnham, 2004; Furnham & Chamorro-Premuzic, 2004; Myszkowski et al., 2014), art expertise (Chamorro-Premuzic & Furnham, 2004; Eysenck & Castle, 1971; Furnham & Chamorro-Premuzic, 2004), or creative potential (Myszkowski et al., 2014).

More generally, this research only aimed at bringing the VAST up to current standards of psychometrical testing through the scope of structural validity. Indeed, from a psychometrical perspective, using a unique

score to measure a construct is justified by a single construct causing most of the variability of the item scores. We did not find such quality with the original VAST (and did not observe an alternate valid structure). We thus revised it in order to obtain such quality. Nevertheless, it should be noted that, as was earlier pointed (Gear, 1986) the very content of the VAST and VAST-R items remains very specific to Götz' art, and it is impossible to conclude from this study that these tests appropriately cover variations of visual aesthetic quality. Also, while revising the VAST, we certainly left aside the most multidimensional items, and this may imply that the construct measured by the VAST-R is different in nature than the construct that Götz aimed at measuring in the first place. In fact, since it is more unidimensional, it may be that the VAST-R is composed of the least artistic items, and measures an ability that is more perceptual in nature than artistic. Nevertheless, this revision of the VAST now captures a unique construct, which can be expected to help future investigations that will attempt to clearly define what this construct is. To this end, research on the relations between the VAST-R and other constructs (art expertise, art practice, intelligence, symmetry recognition, creativity, etc.) may help define its nature.

Also, although model fit estimates indicated appropriate structural validity, we recommend that the structure of the VAST-R be investigated again on other samples. Future studies may notably consider investigating samples with different characteristics (age, gender, aesthetic ability, artistic background, etc.) than that the ones of the convenience sample that was here studied.

Finally, from a purely methodological point of view, the VAST was a typical example of the lack of clarity behind the use of Cronbach's α (Sijtsma, 2009; Zinbarg et al., 2006) in psychometrical investigation processes. Indeed, in Sample 1, a satisfactory α was found, yet all our investigations (McDonald's ω_h, exploratory and confirmatory factor analyses) showed considerable problems in the measure's unidimensionality. While Cronbach's α is heavily used in psychological research, the VAST was a case that shows how it is absolutely not a guarantee of unidimensionality or homogeneity (Sijtsma, 2009).

7. Conclusion

This research is the first one to examine the structural validity of the original VAST. We showed that the VAST has a weak unidimensional structure, but we were able to build a revision, the VAST-R, that presents much better – and satisfactory – structural validity. Such a result was found in both the cross-validation independent sample, and in the overall sample, which we believe provides convincing evidence of this new measure's structural validity.

We recommend the use of the VAST-R as a replacement for the VAST – the list of the 25 original VAST items that form the VAST-R can be found in Table 3. Indeed, the VAST-R appears a suitable measure of visual "T", that has the advantages of being at the same time 1) the shortest measure available, 2) a measure with satisfactory structural validity, 3) an instrument that, although here examined through IRT modeling, offers the practicality of being "securely" scored with sums, and 4) currently the only measure of "T" with satisfactory structural validity.

Table 3
VAST-R items.

Original VAST item	VAST-R item
1	1
2	2
7	3
8	4
9	5
12	6
13	7
15	8
20	9
22	10
23	11
27	12
28	13
31	14
35	15
37	16
38	17
39	18
40	19
41	20
43	21
44	22
45	23
48	24
49	25

References

Bachem, A. (1937). Various types of absolute pitch. *Journal of the Acoustical Society of America*, *9*, 146–151. http://dx.doi.org/10.1121/1.1915919.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1) i-8 10.1002/j.2333-8504.1981.tb01255.x.

Bezruczko, N. (2013). Automatic item generation implemented for measuring artistic judgment aptitude. *Journal of Applied Measurement*, *15*(1), 1–25.

Bezruczko, N., & Frois, J. P. (2011). Comparison of several artistic judgment aptitude dimensions between children in Chicago and Lisbon. *Visual Arts Research*, *37*(1), 1–15.

Bezruczko, N., Manderscheid, E., & Schroeder, D. H. (2016). MRI of an artistic judgment aptitude construct derived from Eysenck's K factor. *Psychology & Neuroscience*, *9*(3), 293–325. http://dx.doi.org/10.1037/pne0000064.

Bezruczko, N., & Schroeder, D. H. (1994). Differences in visual preferences and cognitive aptitudes of professional artists and nonartists. *Empirical Studies of the Arts*, *12*(1), 19–39. http://dx.doi.org/10.2190/92C4-L35B-FC60-89UH.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*.

Bloch, P. H., Brunel, F. F., & Arnold, T. J. (2003). Individual differences in the centrality of visual product aesthetics: Concept and measurement. *Journal of Consumer Research*, *29*(4), 551–565. http://dx.doi.org/10.1086/346250.

Botella, M., Glaveanu, V., Zenasni, F., Storme, M., Myszkowski, N., Wolff, M., & Lubart, T. (2013). How artists create: Creative process and multivariate factors. *Learning and Individual Differences*, *26*, 161–170. http://dx.doi.org/10.1016/j.lindif.2013.02.008.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. http://dx.doi.org/10.1037/h0046743.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. http://dx.doi.org/10.1207/s15327906mbr0102_10.

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

Chamorro-Premuzic, T., & Furnham, A. (2004). Art judgment: A measure related to both personality and intelligence? *Imagination, Cognition and Personality*, *24*(1), 3–24. http://dx.doi.org/10.2190/U4LW-TH9X-80M3-NJ54.

Chan, J., Eysenck, H. J., & Götz, K. O. (1980). A new visual aesthetic sensitivity test: III. Crosscultural comparison between Hong Kong children and adults, and English and Japanese samples. *Perceptual and Motor Skills*, *50*(3, Pt 2), 1325–1326. http://dx.doi.org/10.2466/pms.1980.50.3c.1325.

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts*, *28*(2), 207–222. http://dx.doi.org/10.2190/EM.28.2.f.

Child, I. L. (1964). Observations on the meaning of some measures of esthetic sensitivity. *The Journal of Psychology*, *57*(1), 49–64. http://dx.doi.org/10.1080/00223980.1964.9916671.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. http://dx.doi.org/10.1007/BF02310555.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*(3), 391–404. http://dx.doi.org/10.1207/s15326934crj1803_13.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. http://dx.doi.org/10.1111/bjop.12046.

Enquist, M., & Arak, A. (1994). *Symmetry, beauty and evolution.* 372 (6502), 169–172. Published Online: 10 November 1994; https://doi.org/10.1038/372169a0.

Eysenck, H. J. (1940). The general factor in aesthetic judgements. *British Journal of Psychology. General Section*, *31*(1), 94–102. http://dx.doi.org/10.1111/j.2044-8295.1940.tb00977.x.

Eysenck, H. J. (1967). Factor-analytic study of the Maitland Graves design judgment test. *Perceptual and Motor Skills*, *24*(1), 73–74. http://dx.doi.org/10.2466/pms.1967.24.1.73.

Eysenck, H. J. (1972). Personal preferences, aesthetic sensitivity and personality in trained and untrained subjects. *Journal of Personality*, *40*(4), 544–557. http://dx.doi.org/10.1111/j.1467-6494.1972.tb00079.x.

Eysenck, H. J. (1983). A new measure of "good taste" in visual art. *Leonardo*, *16*(3), 229. http://dx.doi.org/10.2307/1574921.

Eysenck, H. J., & Castle, M. (1971). Comparative study of artists and nonartists on the Maitland Graves design judgment test. *Journal of Applied Psychology*, *55*(4), 389–392. http://dx.doi.org/10.1037/h0031469.

Eysenck, H. J., & Furnham, A. (1993). Personality and the Barron-Welsh art scale. *Perceptual and Motor Skills*, *76*(3), 837–838.

Eysenck, H. J., Götz, K. O., Long, H. Y., Nias, D. K. B., & Ross, M. (1984). A new visual aesthetic sensitivity test: IV. Cross-cultural comparisons between a Chinese sample from Singapore and an English sample. *Personality and Individual Differences*, *5*(5), 599–600. http://dx.doi.org/10.1016/0191-8869(84)90036-9.

Finke, R. A., Ward, T. B., & Smith, S. M. (1996). Creative cognition: Theory, research, and applications. *A Bradford Book*.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299. http://dx.doi.org/10.1037/1040-3590.7.3.286.

Frois, J. P., & Eysenck, H. J. (1995). The visual aesthetic sensitivity test applied to Portuguese children and fine arts students. *Creativity Research Journal*, *8*, 277–284. http://dx.doi.org/10.1207/s15326934crj0803_6.

Furnham, A., & Chamorro-Premuzic, T. (2004). Personality, intelligence, and art. *Personality and Individual Differences*, *36*(3), 705–715. http://dx.doi.org/10.1016/S0191-8869(03)00128-4.

Gear, J. (1986). Eysenck's visual aesthetic sensitivity test (VAST) as an example of the need for explicitness and awareness of context in empirical aesthetics. *Poetics*, *15*(4–6), 555–564. http://dx.doi.org/10.1016/0304-422X(86)90011-2.

Götz, K. O. (1985). *VAST: Visual aesthetic sensitivity test* (4th ed.). Dusseldorf, Germany: Concept Verlag.

Götz, K. O. (1987). Visual aesthetic sensitivity and intelligence. *Perceptual and Motor Skills*, *65*(2), 422. http://dx.doi.org/10.2466/pms.1987.65.2.422.

Götz, K. O., Borisy, A. R., Lynn, R., & Eysenck, H. J. (1979). A new visual aesthetic sensitivity test: I. Construction and psychometric properties. *Perceptual and Motor Skills*, *49*(3), 795–802. http://dx.doi.org/10.2466/pms.1979.49.3.795.

Götz, K. O., & Götz, K. (1974). The Maitland Graves design judgment test judged by 22 experts. *Perceptual and Motor Skills*, *39*(1), 261–262. http://dx.doi.org/10.2466/pms.1974.39.1.261.

Graves, M. E. (1948). *Design judgment test.* New York: Psychological Corporation.

Graves, M. E. (1951). *The art of color and design* (1st ed.). New York, NY: McGraw Hill Book Company, Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: Data mining, inference, and prediction, second edition* (2nd ed.) *2009*. New York, NY: Springer Corr. 7th printing 2013 edition.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205. http://dx.doi.org/10.1177/1094428104263675.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. http://dx.doi.org/10.1007/BF02289447.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

Iwawaki, S., Eysenck, H. J., & Götz, K. O. (1979). A new visual aesthetic sensitivity test: II. Cross-cultural comparison between England and Japan. *Perceptual and Motor Skills*, *49*(3), 859–862. http://dx.doi.org/10.2466/pms.1979.49.3.859.

Jacobsen, T. (2006). Bridging the arts and sciences: A framework for the psychology of aesthetics. *Leonardo*, *39*(2), 155–162.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oaks, Calif: SAGE Publications, Inc.

Kozbelt, A., & Seeley, W. P. (2007). Integrating art historical, psychological, and neuroscientific explanations of artists' advantages in drawing and perception. *Psychology of Aesthetics, Creativity, and the Arts*, *1*(2), 80–90. http://dx.doi.org/10.1037/1931-3896.1.2.80.

Kozbelt, A., Seidel, A., ElBassiouny, A., Mark, Y., & Owen, D. R. (2010). Visual selection contributes to artists' advantages in realistic drawing. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(2), 93–102. http://dx.doi.org/10.1037/a0017657.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(1), 1–26. http://dx.doi.org/10.18637/jss.v028.i05.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (2013 ed.). New York: Springer.

Levitin, D. J., & Rogers, S. E. (2005). Absolute pitch: Perception, coding, and controversies. *Trends in Cognitive Sciences*, *9*(1), 26–33. http://dx.doi.org/10.1016/j.tics.2004.11.007.

McManus, I. C. (2005). Symmetry and asymmetry in aesthetics and the arts. *European Review*, *13*, 157–180. http://dx.doi.org/10.1017/S1062798705000736.

Meier, N. C. (1928). A measure of art talent. *Psychological Monographs*, *39*(2), 184–199. http://dx.doi.org/10.1037/h0093346.

Meier, N. C. (1940). *The Meier art tests: I, art judgment.* Iowa City: Bureau of Educational Research and Service, University of Iowa.

Meier, N. C. (1963). *The Meier art tests: II, aesthetic perception.* Iowa City: Bureau of Educational Research and Service, University of Iowa.

Myszkowski, N., Çelik, P., & Storme, M. (2016a). A meta-analysis of the relationship between intelligence and visual "taste" measures. *Psychology of Aesthetics, Creativity, and the Arts*. http://dx.doi.org/10.1037/aca0000099.

Myszkowski, N., & Storme, M. (2012). How personality traits predict design-driven consumer choices. *Europe's Journal of Psychology*, *8*(4), 641–650. http://dx.doi.org/10.5964/ejop.v8i4.5233.

Myszkowski, N., Storme, M., & Zenasni, F. (2016b). Order in complexity: How Hans Eysenck brought differential psychology and aesthetics together. *Personality and Individual Differences*, *103*, 156–162. http://dx.doi.org/10.1016/j.paid.2016.04.034.

Myszkowski, N., Storme, M., Zenasni, F., & Lubart, T. (2014). Is visual aesthetic sensitivity independent from intelligence, personality and creativity? *Personality and Individual Differences*, *59*, 16–20. http://dx.doi.org/10.1016/j.paid.2013.10.021.

Myszkowski, N., & Zenasni, F. (2016). Individual differences in aesthetic ability: The case for an aesthetic quotient. *Frontiers in Psychology*, *7*(750). http://dx.doi.org/10.3389/fpsyg.2016.00750.

Nodine, C. F., Locher, P. J., & Krupinski, E. A. (1993). The role of formal art training on perception and aesthetic judgment of art compositions. *Leonardo*, *26*(3), 219. http://dx.doi.org/10.2307/1575815.

Revelle, W. (2016). *psych: Procedures for psychological, psychometric, and personality research.* Evanston, Illinois: Northwestern University Retrieved from https://CRAN.R-project.org/package=psych.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. http://dx.doi.org/10.1007/s11336-008-9102-z.

Rosen, J. C. (1955). The Barron-Welsh art scale as a predictor of originality and level of ability among artists. *Journal of Applied Psychology*, *39*(5), 366–367. http://dx.doi.org/10.1037/h0042340.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. http://dx.doi.org/10.1007/s11336-008-9101-0.

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. http://dx.doi.org/10.1037/1931-3896.1.4.247.

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 139–146. http://dx.doi.org/10.1037/1931-3896.2.3.139.

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. http://dx.doi.org/10.2190/EM.27.1.b.

Silvia, P. J., & Nusbaum, E. C. (2011). On personality and piloerection: Individual differences in aesthetic chills and other unusual aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(3), 208–214. http://dx.doi.org/10.1037/a0021914.

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, US: Baywood Publishing Co.

Summerfeldt, L. J., Gilbert, S. J., & Reynolds, M. (2015). Incompleteness, aesthetic sensitivity, and the obsessive-compulsive need for symmetry. *Journal of Behavior Therapy and Experimental Psychiatry*, *49*(Part B), 141–149. http://dx.doi.org/10.1016/j.jbtep.2015.03.006.

Tinio, P. P. L. (2013). From artistic creation to aesthetic reception: The mirror model of art. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(3), 265–275. http://dx.doi.org/10.1037/a0030872.

Ward, T. B. (2007). Creative cognition as a window on creativity. *Methods*, *42*(1), 28–37. http://dx.doi.org/10.1016/j.ymeth.2006.12.002.

Wilson, A., & Chatterjee, A. (2005). The assessment of preference for balance: Introducing a new test. *Empirical Studies of the Arts*, *23*(2), 165–180. http://dx.doi.org/10.2190/B1LR-MVF3-F36X-XR64.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, *21*(1), 299–313. http://dx.doi.org/10.1214/aos/1176349027.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a Scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*(2), 121–144. http://dx.doi.org/10.1177/0146621605278814.