A Psychometric Examination of the Anagram Persistence Task:
More than Two Unsolvable Anagrams May not be Better

Gilles E. Gignac
University of Western Australia
&
Ka Ki Wong
University of Western Australia

A version of which to be published in Assessment (Accepted June 25th 2018)

Author Note
Correspondence should be addressed to Gilles E. Gignac, School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia, 6009, Australia. E-mail: gilles.gignac@uwa.edu.au
Abstract

The purpose of this investigation was to examine a single-, a double-, and multi-anagram versions of the Anagram Persistence Task (APT) for factorial validity, reliability, and convergent validity. Additionally, a battery of intelligence tests was administered to examine convergent validity. Based on an unrestricted factor analysis, two factors were uncovered from the 14 anagram (seven very difficult and seven very easy) response times: test-taking persistence and verbal processing speed. The internal consistency reliabilities for the single-, double-, and multi-anagram (seven difficult anagrams) measures were .42, .85, and .86, respectively. Furthermore, all three versions of the APT correlated positively with intelligence test performance ($r \approx .22$). However, the double-anagram and multi-anagram versions also evidenced negative, non-linear effects with intelligence test performance ($r \approx -.15$), which suggested the possibility of testee adaptation. Taking psychometrics and administration time into consideration, simultaneously, the double-anagram version of the APT may be regarded as preferred.

**Keywords:** persistence, test-taking motivation, testee adaptation
A Psychometric Examination of the Anagram Persistence Task:
More than Two Unsolvable Anagrams May not be Better

For researchers and practitioners alike, the valid interpretation of maximal performance test scores rests, in part, on the assumption that a testee has provided his or her best effort when completing the tests (Revelle, 1987). However, individual differences in test-taking motivation (or persistence) have been reported for people who have completed cognitive-ability type tests, i.e., academic achievement and intelligence tests (Hopfenbeck & Kjaernsli, 2016; Schiel, 1996). Additionally, researchers use measures of persistence to predict the likelihood that addicts will persist to completion programs designed to help them overcome their addictions, for example (Steinberg, et al., 2012).

To-date, several approaches have been developed to measure persistence, including an objective, task-based approach known as the Anagram Persistence Task (Eisenberger & Leonard, 1980). However, little psychometric work has been reported to support the use of the Anagram Persistence Task. Furthermore, various numbers of anagrams have been administered across investigations, from a single essentially unsolvable anagram to as many as 28 combined solvable and essentially unsolvable anagrams. Arguably, the establishment of a brief, simple-to-administer behavioral test of persistence would be attractive, as such a measure would not be affected by the limitations associated with self-report measurement (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Consequently, the purpose of this investigation was to examine psychometrically the scores derived from the Anagram Persistence Task across three methods of scoring: a single essentially unsolvable anagram score, a double essentially unsolvable anagram composite score, and a seven essentially unsolvable anagram composite score.


**Background: Theory and Measures**

Effort was defined by Humphreys and Revelle (1984) as the motivational state of being engaged in, or trying hard at, the completion of a task. Theoretically, test-taking effort has played a central role in the conceptualization and measurement of test-taking motivation. For example, in Humphreys and Revelle’s (1984) model of individual differences in human information processing, arousal and effort were considered key motivational constructs that impact test performance. Additionally, within the Expectancy-Value theory of achievement, the expectancy element of the theory represents the belief that performance is dependent upon effort (Wigfield & Eccles, 2000). Furthermore, Eklöf’s (2010) model of test performance includes test-taking “effort and persistence” (p. 349) as a direct effect. Persistence may be defined as the motive to achieve high standards of performance in the face of frustration (Dudley, Orvis, Lebiecki, & Cortina, 2006). For the purposes of this investigation, we considered test-taking effort and test-taking persistence as representations of the same construct.

Theoretically, it is useful to distinguish between biological intelligence, a theoretical construct defined in abstract terms as an entity’s maximal ability to adapt to the environment successfully that presumably reflects the neurophysiological fitness of the entity, versus psychometric intelligence, the capacity to solve problems on intelligence tests (Eysenck, 1988; Halstead, 1947). A number of non-intellective factors (e.g., competitiveness, compliance with authority), and in particular test-taking motivation, have been theorized and/or shown empirically to be related to psychometric intelligence (Revelle, 1993; Wechsler, 1943; Wise, 2009). In the areas of educational and industrial psychology, a non-negligible amount of research has substantiated a positive association (≈ .20 to .25) between individual differences in test-taking motivation/effort and psychometric intelligence (Duckworth et al., 2011; Gignac, 2018;
Consequently, researchers and practitioners have expressed concerns about the possibility that intelligence test scores may be contaminated by individual differences in test-taking motivation (Wise, 2009). Correspondingly, a small to moderate positive correlation between a measure of test-taking effort/persistence and psychometric intelligence test performance may be regarded as convergent validity for the measure of test-taking effort/persistence.

Typically, in the areas of educational and industrial psychology, researchers measure individual differences in test-taking effort/persistence with self-report measures (Arvey et al., 1990; Thelk et al., 2009). Although self-report measures of test-taking effort have been reported to be associated with some respectable psychometric properties, including internal consistency and factorial validity (Arvey et al., 1990; Thelk et al., 2009), several threats to the criterion validity of self-report measures have been articulated. For example, self-report measures tend to be imbued with socially desirable responding (Gorber, Tremblay, Moher & Gorber, 2007; Podsakoff et al., 2003). Additionally, self-report measures assume that respondents have sufficient introspective ability to generate accurate insights into their behavior (Schmitt, 1994). Finally, scores obtained from self-report response scales are likely to be affected by individual differences in the interpretations of the verbal anchors (Brun & Teigen, 1988; Budescu & Wallsten, 1985). In light of these limitations, objective, behaviourally-based psychometric measures have been recommended, as they do not suffer from the same limitations (Kubinger, 2002; Santacreu, Rubio, & Hernández, 2006). Consequently, research relevant to the establishment of one or more behavioral measures of test-taking effort/persistence may be regarded as useful.
To-date, a small number of behavioral tests of persistence have been used in research. Several of these tests may be categorized as distress tolerance tests, as they measure the length of time people persist in completing an unpleasant task. A relatively commonly administered test in this category is the mirror tracing test, which involves tracing “...the outline of a geometric figure while viewing it through a mirror” (Quinn, Brandon, & Copeland, 1996, p. 187). As the majority of the test shapes are difficult (and frustrating) to trace, individual differences in the amount of time spent on completing the test items is considered a measure of test-taking persistence. Another approach to the measurement of persistence involves measuring the amount of time a participant holds their non-dominant hand and forearm in a container of ice water (Willoughby & Edens, 1996). A similar, complimentary task involves measuring the amount of time a person holds their breadth (Hajek, Belcher, & Stapleton, 1987). As each of these last two approaches are essentially single-item tests, they may not be expected to be associated with a substantial amount of reliability. Additionally, it is arguably not efficient to ensure a container of ice water is available for testing sessions that may involve hundreds of participants.

From a more conventional psychometric perspective, Wise and Kong (2005) suggested that the number of items participants complete on a test very quickly (rapid guessing; e.g., response times less than three seconds for a series of items) may be regarded as an indicator of test-taking effort. Arguably, however, Wise and Kong’s (2005) rapid guessing measurement approach may only be expected to be valid at detecting individuals at the very low-end of the test-taking effort distribution. Additionally, the Wise and Kong (2005) method requires that the researcher determine a minimum amount of time necessary to attend to an item seriously. More recently, Teubner-Rhodes, Vaden, Dubno, and Eckert (2017) developed a regression-based
residual accuracy method from scores derived from the Wisconsin Card Sorting Test – 64 (Kongs et al., 2000). Although the WISCT-64 residual approach shows some promise, it takes between 15 and 20 minutes to administer the WISCT-64 (Kongs, Thompson, Iverson, & Heaton, 2000). Additionally, the regression-based scoring method is somewhat sophisticated, as it requires calculating accuracy residuals on the basis of the correlation between accuracy scores and set-shifting ability scores. In contrast to the measurement methods described above, the Anagram Persistence Task may be regarded as a simple to administer, freely available, conventional psychometric behavioral measure of test-taking persistence.

**Anagram Persistence Task: Administration and Scoring Methods**

Although the use of anagrams in psychometric testing has a long history (Ammons & Ammons, 1959; Miller, 1954), their use in the measurement of persistence was largely initiated by work conducted to test hypotheses relevant to the theory of learned industriousness. In particular, Eisenberger and Leonard (1980) created three lists of 10 anagrams across three levels of difficulty: simple, complex, and unsolvable. The purpose of the administration of the anagrams was simply to train three groups of participants to become accustomed to applying three different levels of effort, rather than measure individual differences, however.

Since Eisenberger and Leonard (1980), several researchers have administered variations of the Anagram Persistence Task to measure individual differences in persistence. For example, Quinn, Brandon and Copeland (1996) administered an eight-anagram version of the Anagram Persistence Task. The first and eighth anagrams were easy to solve (e.g., CEABH = BEACH). By contrast, the second to seventh anagrams (six in total) were essentially unsolvable (e.g., LXYIK = KLYIX). The participants were not made aware that the test included some essentially unsolvable anagrams. Quinn et
al. (1996) stated that such a mixture of items was used to help maintain testee motivation. Testees were given a maximum of 180 seconds to complete each anagram, before being requested to move onto the next anagram. The average amount of time spent on the six unsolvable anagrams was considered a measure of a testee’s test-taking persistence (APT-Difficult subscale). Quinn et al. (1996) found that smokers scored statistically significantly lower than life-long non-smokers on the multi-anagram APT-Difficult subscale ($d = -0.68$). Such a result was interpreted to be consistent with Eisenberg’s (1992) learned industriousness theory (i.e., smokers have a history of continuous reinforcement in the absence of much effort; consequently, they give up relatively easily). Although the results of Quinn et al. (1996) suggest the presence of some criterion-related validity for the multi-anagram version of the Anagram Persistence Task, it should be noted that they did not report any basic psychometric properties associated with the scores (e.g., internal consistency reliability).

In their review of Quinn et al. (1996), Brandon et al. (2003) noted that the order in which the anagrams were administered may not have been optimal. Specifically, all six of the unsolvable anagrams were presented successively (i.e., anagrams two to seven). Consequently, Brandon et al. (2003) suggested that the participants in Quinn et al. (1996) may have suspected that the anagrams were unsolvable, prior to completing all of the anagrams, which may have affected their behavior, independently of their natural inclination to persist. Therefore, as an extension to Quinn et al. (1996), Brandon et al. (2003) administered an 11 anagram version of the Anagram Persistence Task to a sample of 144 smokers enrolled in smoking cessation program. Five of the anagrams were relatively easy to solve (CEABH = BEACH) and the remaining six anagrams were essentially unsolvable (e.g., LXYIK = KYLIX). In contrast to Quinn et al. (1996), the easy and unsolvable anagrams were presented in a more mixed sequence.
Specifically, items 2, 3, 4, 6, 7, and 8 were unsolvable anagrams and items 1, 5, 9, 10, 11 were easy anagrams. Furthermore, the testees were given a maximum of 240 seconds to complete any particular anagram, after which the next anagram was presented automatically. The mean amount of time spent on the six essentially unsolvable anagrams was considered a composite measure of test-taking persistence.

As a rare example in the literature, Brandon et al. (2003) reported a coefficient alpha of .85 for the APT-Difficult subscale scores. However, contrary to what was hypothesized, the APT-Difficult subscale scores failed to predict sustained abstinence from smoking, one year after the completion of the program. Thus, the results of Brandon et al. (2003) suggest that the multi-anagram composite Anagram Persistence Task scores may be associated with respectable internal consistency reliability. However, the validity of the multi-anagram persistence composite scores, based on the sum of response times to six unsolvable anagrams, may be questioned.

More recently, in an investigation relevant to video-gamers, Ventura, Shute and Zhao (2013) administered a 28-item version of the Anagram Persistence Task: 12 essentially unsolvable anagrams and 16 easy anagrams. The difficult anagrams were selected from a dictionary which included obscure words (e.g., digamy, gowpen). As per Brandon et al. (2003), the essentially unsolvable and easy anagrams were presented in a mixed order to help ensure the participants did not receive too many difficult items, consecutively. Furthermore, the participants were given a maximum of 120 seconds to solve each anagram. After 120 seconds, the next item was presented, automatically. Based on a sample of 102 undergraduate students, Ventura et al. (2013) found that the unsolvable and easy anagrams were associated with solve-rates of 24% and 95%, respectively. Additionally, the 12 unsolvable anagram APT-Difficult subscale correlated positively with self-reported conscientiousness ($r = .27$), which suggested
some level of convergent validity. However, consistent with virtually all other investigations that used an anagram task for the measurement of persistence/test-taking motivation, Ventura et al. (2013) did not report any psychometric results for the multi-anagram APT-Difficult subscale scores (e.g., factorial validity; internal consistency reliability).

In contrast to the multi-anagram versions of the Anagram Persistence Task, some researchers have created persistence composite scores, based on the sum of the response times to only two unsolvable anagrams. The impetus for not using a multi-item Anagram Persistence Task relates to concerns that testees discern the purpose of the testing after two or three essentially unsolvable anagrams and alter their behavior, accordingly. For example, in a study relevant to automobile driver anger, Cackowski and Nasar (2003) administered an Anagram Persistence Task with four items: two easy anagrams (IUFTR = FRUIT; OEWRP = POWER) and two essentially unsolvable anagrams (ANBIT; DATGI). The order of anagram item administration was the following: easy, unsolvable, unsolvable, and easy. Participants were asked to move onto the next anagram, if they did not do so within 300 seconds. A composite score was created, based on the sum of time spent on the two unsolvable anagrams (double-anagram APT-Difficult subscale). Cackowski and Nasar (2003) found that participants who viewed a video with a scenic parkway evidenced greater persistence on the double-anagram APT-Difficult subscale, in comparison to participants who viewed a video of a road through a built-up area. Thus, the result was consistent with their theoretical prediction. However, no psychometric evidence was reported for the double-anagram APT subscale scores. Dejoy (1985) used the same double-anagram APT subscale scores in a study relevant to tolerance to noise. As per Cackowski and Nasar (2003), some
theoretically consistent results were reported. However, no psychometric evidence was reported supporting the use of double-anagram subscale scores.

Finally, in contrast to the multi-anagram and double-anagram versions of the Anagram Persistence Task, some researchers have administered a version of the Anagram Persistence Task that included only one unsolvable anagram (Carver, Peterson, Follansbee, & Scheier, 1983; Solberg Nes, Segerstrom, Sephton, 2005; Solberg Nes, Carlson, Crofford, de Leeuw, & Segerstrom, 2011). In these studies, the participants were given 20 minutes to solve a total of 11 anagrams. However, the first anagram was unsolvable (GGAWIL). If, after 300 seconds, the participant continued to persist with the first anagram, the participant was instructed to move onto the second anagram. The amount of time spent on the first anagram was considered an indicator of test-taking persistence. Perhaps surprisingly, theoretically consistent results have been reported, based on the single-anagram version of the Anagram Persistence Task. For example, Soldberge Nes et al. (2005) found that optimism was related positively to the amount of time spent on a single unsolvable anagram. Furthermore, the effect of optimism interacted with public self-consciousness. However, the results with the single-anagram APT-Difficult score have not be consistent. For example, Soldberge Nes et al. (2011) failed to observe a statistically significant correlation between self-reported conscientiousness and time spent on the first unsolvable anagram.

As noted above, some researchers have used single and double Anagram Persistence Task scores, due to concerns that the testees may discern the purpose of the testing, when exposed to multiple unsolvable anagrams (say, three or more). Consequently, only the first one or two unsolvable anagrams are implied to be relatively pure indicators of test-taking persistence. However, there is little psychometric research to support the notion that testees alter their behavior across unsolvable anagrams.
Furthermore, it is has been commonly argued that single-item measures cannot be expected to yield sufficiently reliable test scores (Loo, 2002; McIver & Carmines, 1981). Consequently, greater levels of reliability and validity may be observed (e.g., correlations with cognitive ability type test performance), based on a multi-anagram version of the Anagram Persistence Task, in comparison to the single-anagram and double-anagram versions. Ultimately, little psychometric work has been published to help support the use of any version of the Anagram Persistence Task. Validation work relevant to the Anagram Persistence Task may be considered worthwhile, as it is a simple test to administer, free of charge, and may take less than five minutes to administer, depending on the number of difficult anagrams that are administered (i.e., < 3). Based on theoretical and empirical considerations (Duckworth et al., 2011; Revelle, 1993), a small to moderate positive correlation between the APT and psychometric intelligence test performance would be regarded as convergent validity for the APT.

**Summary & Purpose**

A review of the literature suggests that researchers who use the Anagram Persistence Task tend to administer a mixture of solvable and essentially unsolvable anagrams. However, various researchers have calculated test-taking persistence, based on different numbers of unsolvable anagrams (one, two, or more), and it is unknown which method may be preferred psychometrically. Arguably, the establishment of a relatively brief, accessible, and valid behavioural measure of test-taking persistence would be valuable, as the measurement of test-taking persistence has been observed across a variety of psychological disciplines, including addiction (e.g., Brandon et al., 2003; Quinn et al., 1996), psychological therapy (e.g., Rodman, Daughters, & Lejuez, 2009), job recruitment (e.g., Arvey, Stickland, Drauden, & Martin, 1990; Chan,
Schmitt, DeShon, Clause, & Delbridge, 1997), and clinical neuropsychology (e.g., Busse & Whiteside, 2012). Consequently, the purpose of this investigation was to evaluate test scores from a 14-item version of the Anagram Persistence Task (seven easy and seven essentially unsolvable anagrams) for factorial validity and internal consistency reliability, across three methods of scoring: single-item, double-anagram, and multiple-anagram. Additionally, convergent validity for the three scoring methods derivable from the Anagram Persistence Task was evaluated, by estimating the association between the three Anagram Persistence Task scores and performance on a small battery of intelligence tests.

Method

Sample

The sample consisted of 173 first-year undergraduate university students (68.8% female). Such a sample size was associated with statistical power of .73 to detect a correlation of |.20| as statistically significant ($p < .05$). The participants were recruited from a first-year undergraduate psychology research pool within a large, English speaking university in Australia. The students participated in the research voluntarily for a small amount of extra course credit. Median age was 18 (age range: 17 to 43 years; $M = 20.00; SD = 4.75$). Although information on ethnicity was not obtained from the participants, the university student body is known to be populated from a primarily white, European background. Recruits were required to speak English as a first language to participate. Research ethics approval was obtained for this investigation from the university’s committee.

Measures

Test-taking persistence. Test-taking persistence was measured with a version of the 28-item Anagram Persistence Task (APT) employed by Ventura et al. (2013).
Specifically, in this investigation, seven easy anagrams and seven difficult anagrams (14 total) were selected from the anagrams listed in Appendix A of Ventura et al. (2013). For the purposes of test administration, the easy anagrams were presented in the following item-trial order: 1 (lalb = ball), 2 (norb = born), 4 (yaaw = away), 6 (bluc = club), 9 (leph = help), 10 (moeh = home), and 12 (fels = self). The difficult (i.e., essentially unsolvable) anagrams were presented in the following item-trial order: 3 (diyamg = digamy), 5 (wopneg = gowpen), 7 (hamyap = mayhap), 8 (cytepl = yclept), 11 (styrt = tryst), 13 (libee = belie), and 14 (hethy = hythe). Thus, the easy and difficult anagrams were semi-randomly mixed. Correspondingly, the easy and difficult anagrams had roughly comparable item-order presentation means of 6.3 and 8.7, respectively.\(^1\) Prior to starting the APT, the participants were provided with the following information on the screen (as per Ventura et al., 2013):

“In this task you will be asked to solve 14 anagrams. Anagrams are a string of letters that can be reordered to make a word. For example, YBO would be an anagram for BOY. Your task is to rearrange the letters into a word and type in your guess. Some words will be harder than others, but try to do your best to make a word out of the anagram. All of the anagrams are homemade, so you will not find any answers online. If you are having too much trouble with an anagram, you can skip it by clicking on the ‘Next’ button.”

After 120 seconds, the next anagram was presented on the screen, automatically (“Okay, try this next one”). Time in seconds was the unit of measurement in this investigation. For thoroughness, we note that mean solve rates for the APT-Easy and APT-Difficult anagram composites (7 anagrams each) were 94.7% and 11.2%, respectively.

---

\(^1\)For example, APT-Easy = (1 + 2 + 4 + 6 + 9 + 10 + 12)/7 = 6.3.
**Intelligence.** Consistent with the Wechsler scales which measure four index scores of cognitive ability, intelligence was measured in this investigation with four corresponding diverse tests of intelligence. *Crystallized intelligence* was measured with the Advanced Vocabulary Test (AVT; Gignac, Shankaralingam, Walker & Kilpatrick, 2016). *Fluid intelligence* was measured with the odd items from the Advanced Progressive Matrices (APM; Raven, 1998). Participants were given 10 minutes to complete the APM short-form. *Memory span* was measured with a slight adaptation of the Letter-Number Sequencing test (LNS; Wechsler, 2008). *Processing speed* was measured with the four subtests within the Connections battery (Salthouse et al., 2000). Two of the Connections subtests are similar to the well-known Trails A test and two are similar to Trails B (see Reitan, 1958). As a general factor of intelligence has been contended to be best measured from a correlation matrix that is balanced, i.e., roughly equal numbers of indicators from each sub-dimension (see, for example, Ashton, Lee., & Vernon, 2001), it was considered useful to reduce the four processing speed tasks cores into a single processing speed composite score (via principal component analysis). Norm referenced scoring was not available for the intelligence tests. However, a negative correlation was observed between age and only one of the intelligence tests (i.e., Connections component scores \( r = -.20, p = .007 \)). Consequently, the processing speed component scores were residualised of their effects due to age. All analyses were performed with the residualised processing speed component scores.

Finally, the four intelligence test scores were submitted to a principal component analysis to obtain a representation of general intelligence. Correspondingly, a parallel analysis (O’Connor, 2000; 3000 re-samples; 95th percentile) suggested the extraction of a single-component (46.9% of the variance), as only a single eigenvalue exceeded
the rank-ordered eigenvalues derived from random data (raw data eigenvalues = 1.88, .85, .69, .58; random data eigenvalues = 1.28, 1.11, 1.00, .92). The general intelligence component loadings were: AVT = .58, APM = .72, LNS = .66, speed = .77. Component scores (regression-based) were estimated for each participant to represent general intelligence test performance. According to Armor’s (1973) theta, the internal consistency reliability associated with the general component scores was .63. Where appropriate, the correlations with the general intelligence component scores were disattenuated for imperfect reliability with the classical test theory approach ($r_c$; Nunnally & Bernstein, 1994). Component scores, rather than factor scores, were used, here, as they do not suffer from factor indeterminacy (Velicer, 1976) and internal consistency reliability can be calculated for component scores (Armor, 1973).

**Procedure**

The test scores examined in this investigation were derived from a larger study relevant to intelligence and the sexual appeal of intelligence. All participants were tested individually, and the testing sessions lasted approximately 50 minutes in total. After reading an information sheet, the participants signed an online consent form and completed basic demographic questions. Then, the participants completed several self-report questionnaires ($\approx$ 12 minutes) on a computer. Next, the participants completed the Anagram Persistence Task on a computer. Finally, the participants completed the intelligence testing in the following order: Advanced Vocabulary Test (computer), Advanced Progressive Matrices (computer), Connections test (face-to-face) and Letter-Number Sequencing (face-to-face). A test administrator was in the testing room at all times.

**Data Analysis**
All analyses were conducted with SPSS. In order to evaluate the factorial validity of the Anagram Persistence Task scores, the 14 anagram response times were submitted to a principal-axis factor analysis with promax (oblique) rotation. A raw data permutation-based parallel analysis (95th percentiles) was performed to determine the number of factors to extract (O’Connor, 2000). Standardized pattern matrix loadings greater than |.20| were considered salient.

Consistent with Wanous and Hudy (2001), we estimated the reliability associated with the single-anagram APT-Difficult scores via its communality within the relevant factor analytic solution. We estimated the internal consistency reliability associated with the double-anagram APT-Difficult scores via the Spearman-Brown formula, as recommended by Eisinga, Te Grotenhuis, and Pelzer’s (2013). Finally, we estimated the internal consistency reliability associated with the multi-anagram APT-Difficult scores via the well-known coefficient alpha.

In order to test the hypothesis that testees changed their behaviour across unsolvable anagrams (i.e., become progressively demotivated), we conducted a factorial within-subjects ANOVA with anagram difficulty type as a within-subjects factor (two levels: easy vs. difficult) and item-order as a within-subjects factor (i.e., seven levels: anagram presentation 1 to 7). Although no specific hypothesis was considered for the APT-Easy anagrams, it was, nonetheless, considered useful to compare the behaviour of the participants across roughly comparably presented easy and unsolvable anagrams.

The associations between the various APT-Difficult scale scores (single-anagram, double-anagram, multi-anagram) and general intelligence were estimated via Pearson correlations. In the event that non-linear effects were suspected in the scatter plots, a series of hierarchical multiple regression analyses were performed (Pedhazur, 1997). In
the first step, the linear effect was estimated. In the second step, the non-linear effect was estimated, whereby the APT-Difficult scale scores were squared to represent a quadratic effect. The observation of a statistically significant beta-weight associated with the quadratic effect term was considered statistical evidence in favor of a non-linear effect.

It will be noted that an examination of the distribution of the Anagram Persistence Test item scores uncovered a number of outliers for several of the APT-Difficult anagrams, based on the outlier inter-quartile range rule with a 3.0 multiplier (Hoaglin & Iglewicz, 1987). However, there was no compelling reason to question the validity of the outlying scores. Consequently, the outlying values were Winsorized (reduced in magnitude to the next largest data point in the distribution not suspected to be an outlier; Tukey, 1962). Some of the Winsorized APT-Difficult anagrams were, nonetheless, associated with an appreciable level of positive skew (e.g., item 10: skew = 1.72). Consequently, the standard errors, p-values, and 95% confidence intervals were estimated via bootstrapping (2000 resamples; bias-corrected confidence intervals), where possible.

**Results**

**Factor Structure and Internal Consistency**

As can be seen in Table 1, the Anagram Persistence Task inter-anagram correlations tended to be positive. The mean inter-anagram correlation associated with the APT-Difficult anagrams was .46 (range .20 to .73). By comparison, the mean inter-anagram correlation associated with the APT-Easy anagrams was .12 (range -.02 to .29). Thus, there was a numerically greater amount of common variance associated with the APT-Difficult anagrams. Finally, the KMO measure of sampling adequacy was
.78, which suggested that there was a sufficient amount of shared variance between the anagrams to merit a factor analysis (i.e., >.70; Kaiser, 1974).

The parallel analysis suggested the extraction of three factors. However, a three factor solution was not considered interpretable (i.e., the third factor was defined by only two items). Consequently, the principal-axis factor analysis was re-conducted with the extraction of two factors. The two factor solution accounted for 32.2% of the variance. As can be seen in Table 2, the first factor was defined by the APT-Difficult anagrams (sum of squared loadings = 3.44; 24.6% of the variance) and was considered representative of persistence. The second factor was defined by the APT-Easy anagrams (sum of squared loadings = 1.10; 7.56% of the variance) and was considered representative of processing speed. The factor solution was associated with appreciable simple structure, as only one item had a cross-loading in excess of |.20| (i.e., item 11). The correlation between the factors was estimated at $r = .06$. Thus, the persistence and processing speed dimensions were essentially unrelated. The residual correlation matrix was associated with a mean absolute value of only .04, which suggested that the extraction of a third factor was contra-indicated. However, one appreciably large residual correlation between anagrams 13 and 14 (both from the APT-Difficult scale) was noted ($r = .33$).

As can be seen in Table 3 (right side), the double-anagram and multi-anagram APT-Difficult scale scores were associated with respectable levels of internal consistency reliability (.85 to .91). By contrast, the single-anagram APT-Difficult score was associated with a reliability of .42, based on its communality associated with the two-factor solution. Finally, for thoroughness, we note that the APT-Easy subscale scores were associated with a coefficient alpha of .41.

**Mean Differences: Main and Interaction Effects**
To test the differences between the APT-Easy and APT-Difficult mean anagram response times, a 2 x 7 factorial repeated measures ANOVA was conducted, with anagram difficulty as a within-subjects factor with two levels (i.e., easy vs. difficult) and anagram order as a within-subjects factor with seven levels (i.e., item presentation 1 to 7). As the assumption of sphericity was violated for the item order within-subjects factor, $\chi^2(20) = 176.32, p < .001$, Huynh-Feldt $\varepsilon = .74$, the Huynh-Feldt adjusted results were consulted and reported.

As expected, the main effect of anagram type (easy vs. difficult) was statistically significant, $F(1, 172) = 568.76, p < .001$, partial $\eta^2 = .768$. On average, the participants spent less time on the APT-Easy anagrams ($M = 9.09, SD = 2.45$), in comparison to the APT-Difficult anagrams ($M = 49.18, SD = 22.39$). Additionally, the main effect of anagram order was also statistically significant, $F = 64.28, p < .001$, partial $\eta^2 = .272$. However, there was a statistically significant interaction between the anagram difficulty and anagram order within-subjects factors, $F(4.51, 775.20) = 53.96, p < .001$, partial $\eta^2 = .239$. As can be seen in Figure 1 (Panel A), the APT-Difficult anagram mean response times (in seconds) decreased across anagram presentation order, whereas the APT-Easy anagram mean response times remained relatively stable across anagram order. A series of follow-up paired samples $t$-tests revealed that all seven APT-Easy vs. APT-Difficult mean response time comparisons were statistically significant ($p < .001$). Thus, for each corresponding easy and difficult anagram, the participants spent more time completing the difficult anagrams, as expected. However, as can be discerned by an examination of Figure 1, the statistically significant interaction implied that the magnitude of the difference decreased statistically significantly across anagram order: comparison 1, $d = -2.91$; comparison 2, $d = -2.81$; comparison 3, $d = -2.15$; comparison 4, $d = -1.97$; comparison 5, $d = -1.71$;
comparison 6, \( d = -2.53 \); comparison 7, \( d = -1.74 \). As some of the essentially unsolvable anagrams were solved correctly, the analysis was re-run on the mean reaction times that excluded the correctly solved APT-Difficult reaction times. As can be seen in Figure 1 (panel B), the pattern of effects was very similar.

**Associations with Intelligence**

Consistent with the preceding analysis, the three types of ATP-Difficult response time scores (single-, double-, and multi-anagram) were calculated on the basis of difficult anagrams that were not solved accurately by the participant to better reflect participant’s test-taking effort. Furthermore, for thoroughness, all seven contiguous item composite score combinations were correlated with intelligence test performance. As can be seen in Table 3, all seven types of Anagram Persistence Test scores (single, double, multi-items) correlated positively and statistically significantly with the general intelligence test scores. Furthermore, the correlations were essentially identical in magnitude across all types of composite scores (\( r = .19 \) to .22).

However, as can be seen in Figure 2, the corresponding scatter plots suggested the possibility of a curvilinear effect between the APT-Difficult scores and intelligence test performance. For conciseness, only the results associated with the seven-anagram, double-anagram, and single-anagram versions of the APT-Difficult are reported. Based on a series of curvilinear hierarchical multiple regressions, there was some statistical evidence to suggest that the multi-anagram and the double-anagram APT-Difficult scales were associated with inverted U-shaped effects with general intelligence test scores, on the basis of the negative quadratic function beta weights and corresponding semi-partial correlations (multi-anagram: \( \beta = -.11 \), 95\%: -.26/.01, semi-partial \( r = -.12 \), semi-partial \( r_c = -.15 \); double-anagram: \( \beta = -.19 \), 95\%: -.35/-01, semi-partial \( r = -.16 \), semi-partial \( r_c = -.20 \)). By contrast, there was no evidence to suggest a curvilinear
effect between the single-anagram APT-Difficult scores and general intelligence test scores ($\beta = -.05$, 95%CI: -.24/.14, semi-partial $r = -.04$, semi-partial $r_c = -.05$).

Finally, the seven APT-Difficult anagram response times were correlated with the general intelligence test scores, individually (i.e., at the item-level). As can be seen in Table 4 (middle), there was inconsistency in the results. Specifically, only anagrams 3, 5 and 14 correlated positively and statistically significantly with the general intelligence test scores. By contrast, anagram 11 correlated negatively with the general intelligence test scores, $r = -.17$, $p = .025$. However, excluding correctly solved anagram reaction times, anagram 11 was no longer observed to correlate negatively with intelligence test performance ($r = -.01$).

Next, for the purposes of thoroughness, the APT-Easy anagram data were analysed. The association between the APT-Easy scale scores and general intelligence was estimated at $r = -.35$, $p < .001$, 95%CI: -.50/-21. Thus, faster response times to the easy anagrams tended to be associated with higher levels of general intelligence. As can be seen in Figure 2 (Panel D), there was an absence of visually apparent curvilinear effect, which was verified by the absence of a statistically significant quadratic term in a hierarchical multiple regression analysis ($\beta = .01$, $p = .987$, 95%CI: -.10/.21, semi-partial $r = .01$). Again, for thoroughness, the seven APT-Easy anagram response times were correlated with general intelligence test performance, individually. As can be seen in Table 4 (left-side), the APT-Easy anagrams tended to correlate negatively with general intelligence test performance (e.g., item 1: $r = -.23$, $p = .001$).

**Discussion**

Distinct APT-Difficult (persistence) and APT-Easy (processing speed) response time factors were identified in this investigation. Additionally, the double-anagram and multi-anagram APT-Difficult scale scores were associated with respectable internal
consistency reliabilities. Furthermore, all versions of the APT-Difficult scales (single, double, and multi) were associated with general intelligence test performance positively. However, there was also evidence to suggest that the double- and multi-anagram APT scores related to general intelligence in a negative, curvilinear fashion. These key results are discussed in detail below.

**Factor Structure and Reliability**

The results of the factor analysis suggested the presence of two interpretable factors, consistent with test-taking persistence (APT-Difficult anagrams) and verbal processing speed (APT-Easy anagrams). The persistence and verbal processing speed factors unlikely arose because of differences in item means (i.e., difficulty factors), as the item data were not dichotomous in nature (McDonald & Ahlawat, 1974). Consequently, the multi-item Anagram Persistence Task used in this investigation may be considered to be associated with respectable factorial validity.

Although the APT-Difficult anagram factor was relatively strong (loadings > .80), we note that the later administered anagrams (11, 13, and 14) tended to be associated with less strong loadings (∼ .55), in comparison to the earlier presented (3, 5, 7) APT-Difficult anagrams (∼ .80). Such a pattern of results lends some support to previously expressed concerns about the validity of response times to later administered essentially unsolvable anagrams (e.g., Carver et al., 1983; Cackowski & Nasar, 2003), a topic we revisit further below.

The internal consistency reliability associated with the multi-anagram APT-Difficult scale scores (i.e., $\alpha \approx .85$ to .90) was essentially identical to the estimate reported by Brandon et al. (2003) for their 6-anagram APT-Difficult scale ($\alpha = .85$). Thus, it may be contended that a multi-anagram version of the Anagram Persistence Task (≥ 3 anagrams) may yield scores sufficiently reliable for research purposes and
possibly even clinical applications (Nunnally & Bernstein, 1994). Perhaps surprisingly, the internal consistency reliability associated with the double-anagram version of the APT-Difficult scale (i.e., $\alpha = .85$) was essentially equal to the multi-anagram APT-Difficult scales. Thus, with respect to internal consistency reliability, there does not appear to be any benefit to administering more than two unsolvable anagrams. Such a result is consistent with the observation of some statistically significant convergent/predictive validity for the double-anagram version of the APT-Difficult subscale (e.g., Cackowski & Nasar, 2003; Dejoy, 1985).

In contrast to the multi-anagram and double-anagram APT-Difficult subscales, the internal consistency reliability associated with the single-anagram APT-Difficult scores was estimated to be unacceptably low for researcher purposes (i.e., $\alpha = .42$; Nunnally & Bernstein, 1994). Given the positive association between number of items and reliability (Cortina, 1993), such a result should not be surprising. The estimated reliability of .42 may have been depressed meaningfully, due to the substantial ceiling effect associated with the single-anagram scores. As can be seen in Figure 2 (Panel C), a substantial percentage of the participants ($\approx 20\%$) reached the maximum 120 second mark on the first unsolvable anagram. Due to test administration time limitations, a maximum of 120 seconds was allowed for each anagram. It is plausible to suggest that a longer time allowance (e.g., 180 or 240 seconds) may yield greater variability in the scores and, consequently, greater reliability for the single-anagram version of the APT-Difficult scale. Single-item scales with respectable reliabilities have been reported (Bergkvist, 2015; Nichols & Webster, 2013; Postmes, Haslam, & Jans, 2013). Furthermore, with enhanced internal consistency reliability for the single-anagram scale of persistence, greater validity may also be expected. Further research to evaluate the
optimal time allowance for the administration of the single-anagram version of the Anagram Persistence Test is encouraged.

In contrast to the APT-Difficult factor, the APT-Easy factor was relatively weak. Perhaps the main reason for such an observation was that the standard deviations associated with the APT-Easy anagrams were multiple times smaller than the APT-Difficult anagram standard deviations (see Table 1; right-side). Thus, as variability is an inherently valuable quality for test scores (Nunnally & Bernstein, 1994), the relatively small amount of variability associated with the APT-Easy anagram response times may have precluded the possibility of substantial inter-anagram correlations. Consequently, although not the purpose of this investigation, the APT-Easy subscale scores obtained in this investigation should probably not be viewed as a psychometrically robust indicator of verbal processing speed.

As we believe this investigation is the first factor analysis associated with any version of the Anagram Persistence Task, the factor analysis results are difficult to integrate within the persistence measurement literature. Furthermore, in the reaction time research area more broadly, we note that researchers tend to factor analyze the data at the subscale level, rather than the item-level (e.g., Kranzler & Jensen, 1991; Miller & Vernon, 1996). We, therefore, encourage more item-level factor analytic work in the broader area of individual differences in response/reaction times. Such work may help uncover the reasons for the relatively poor psychometrics associated with commonly administered tasks based on reaction times (e.g., stroop, flanker; see Hedge, Powell, Summer, 2017).

Lastly, we note briefly that the internal consistency reliability associated with the APT-Easy subscale scores was unacceptably low at .41. Admittedly, researchers do not use the Anagram Persistence Task to measure individual differences in verbal
processing speed. However, based on the mean APT-Easy anagram response time inter-correlation of .12, the Spearman-Brown prophecy formula suggests that approximately 20 easy anagrams would yield a composite score with sufficient reliability (i.e., > .70) for exploratory research.

**Mean Item Persistence Times**

Consistent with the pattern of reduced factorial validity associated with each progressively administered APT-Difficult anagram, there was an observed reduction in anagram mean persistence times. Additionally, the essentially linear reduction in anagram mean persistence times are consistent with the pattern of unsolved anagram mean persistence times reported by Ventura et al. (2013). It should be made clear, however, that the group-level mean response time reductions across unsolvable anagrams should not be considered necessarily a threat to validity, as individual differences in test-taking persistence is the focus of the assessment. Instead, in this context, the concern is that a non-negligible percentage of the participants may discern the distinction between unsolvable and solvable anagrams, as the administration of the test progresses.

Such a possibility is a realistic one. Consider, for example, that the APT-Easy anagrams were solved in this investigation, consistently, within approximately 10 seconds, on average. By contrast, the participants spent up to 120 seconds on the APT-Difficult anagrams. Consequently, after the presentation of the first few easy and difficult anagrams, a non-negligible percentage of the participants may have discerned that, if an anagram were not solved within approximately 10 seconds, it was unsolvable. Unfortunately, we did not ask the participants their impressions, after completing the APT. Arguably, discernment between the two types of anagrams may be considered an indicator of intelligence, rather than persistence. Importantly, in contrast
to the multi-anagram version of the Anagram Persistence Task, exposure to only one or two essentially unsolvable anagrams does not afford the testee the realistic opportunity to form an impression of the two types of items within the scale. Thus, previously expressed validity related concerns about administering more than one or two unsolvable anagrams to testees (e.g., Carver et al., 1983; Cackowski & Nasar, 2003) appear to be valid, a topic we discuss in further detail below.

**Convergent Validity**

Theoretically, the amount of effort applied to complete cognitive ability type tests has been hypothesized to influence psychometric intelligence test performance (Revelle, 1987). The hypothesis was supported in this investigation, as the amount of time spent completing unsolvable anagrams was found to relate positively to general intelligence test scores at $r = .19$ to $.22$ ($r_c = .24$ to .28). The magnitude of the effect size may be regarded as relatively typical in individual differences research (Gignac & Szodorai, 2016). Furthermore, the effect size was comparable in magnitude to the effects reported for self-reported measures of test-taking motivation (Fervaha et al., 2011; Thelk et al., 2009). Thus, the possibility that a behavioral measure would evidence greater validity than a self-report measure was not suggested by this investigation. It should be noted, however, that the APT was administered prior to the intelligence tests. By contrast, self-report measures of test-taking motivation are typically administered immediately after the intelligence testing. Therefore, as test-taking motivation is known to change throughout a testing session (Penk & Richter, 2017), more substantial effect sizes may be observed, when the APT is administered in the middle of an intelligence testing session. Further research is encouraged, here.

At least superficially, the positive correlation between the APT-Difficult subscale and general intelligence test performance supports contentions that individual
differences in test-taking motivation may affect the validity of psychometric intelligence test scores negatively (e.g., Duckworth et al., 2011; Wise, 2009). However, the direction of the effect between test-taking effort and test performance in adults is mostly only assumed to flow from test-taking effort to intelligence. It is plausible to suggest that intelligence may influence test-taking motivation - at least in samples of adults who are at least moderately motivated to complete the testing. Further research with latent variable deficit and interference models may help determine the more plausible direction of the effect between the two dimensions (see Halpin, da-Silva, & De Boeck, 2014). In the area of test-taking anxiety, for example, the balance of the latent variable evidence suggests that ability influences test-taking anxiety, rather than the other way around (Sommer & Arendasy, 2015; Sommer & Arendasy, 2016). Based on the curvilinear regression analyses, the association between the double-anagram and the multi-anagram APT-Difficult scores and general intelligence was negative in direction at the highest spectrum of the APT-Difficult reaction time scores. Stated alternatively, the tendency to persist with attempts to complete the later administered, essentially unsolvable anagrams suggested the presence of relatively less general intelligence. Such a result is consistent with the possibility that more intelligent people adapt to the Anagram Persistence Task, by ceasing to persist beyond a certain point. Correspondingly, it has been suggested previously that not all persistence is adaptive (Janoff-Bulman & Brickman, 1982). That is, in the context of the multi-anagram version of the Anagram Persistence Task, it is arguably adaptive to discern that some of the anagrams may be impossible to solve and to move onto those that are solvable. To our knowledge, this is the first investigation to examine the possibility of curvilinear effects between test-taking persistence and a criterion variable. Consequently, further research on the possible construct of ‘adaptive persistence’ (or
'excessive persistence', see Baumeister & Heatherton, 1996) is encouraged with the multi-anagram version of the Anagram Persistence Task. By contrast, the double-anagram version of the Anagram Persistence Task may be considered the preferred version of the Anagram Persistence Task to assess persistence, considering both psychometrics and test administration time, simultaneously. However, a single-anagram task, with a maximum time to respond closer to 180 seconds, may be shown to be the ideal, in a future investigation. With respect to multi-anagram versions of the APT, it would be useful to administer easy and difficult anagrams with the same number of letters to eliminate the possibility that some testees discern the difference between the two types of items on such a simple basis.

It is noteworthy that not all of the APT-Difficult scale items exhibited correct response rates consistent with the notion of being essentially unsolvable. In particular, 43.4% of the participants completed item 11 (styrt = tryst) correctly. Including all reaction times in the analysis, item 11 correlated negatively with general intelligence test performance ($r = -.17$). As a result, the anagram styrt behaved, to some degree, in a manner similar to the APT-Easy anagrams. Specifically, a non-negligible percentage of the testees appear to have submitted their response relatively quickly and moved onto the next anagram. Such a result suggests that users of an Anagram Persistence Task cannot necessarily assume that, simply because an anagram is based on a very rare word, it will behave entirely like an essentially unsolvable anagram. To-date, researchers who have used the APT tend to calculate composite reaction times, without taking into consideration correct response rates. Although researchers can calculate mean APT response times that exclude correctly solved difficult anagrams, as performed in this investigation, ideally, essentially unsolvable anagrams would have correct response rates close to zero percent to help ensure the validity of the test, as
well as the need to calculate composite scores on different numbers of anagrams across participants. On this basis, we underscore the value of conducting item-level analyses for a scale – a unique feature of this investigation with respect to the APT.

**Limitations**

Although the sample size used in this investigation was reasonably acceptable for a psychometric evaluation of small scale (< 15 items), it should be acknowledged that replication with a larger sample size would be useful. In particular, greater confidence in the correlations between the individual anagram response times and intelligence test performance would be afforded by an analysis based on a sample size closer to 500. Additionally, a number of statistical analyses were performed in this investigation, without making any adjustments to familywise error, again, underscoring the benefits of replication.

A larger sample would also afford the opportunity to evaluate the association between persistence and a criterion variable (e.g., intelligence) with more sophisticated analyses, such as a latent variable curvilinear modeling (Marsh, Wen, & Hau, 2006). Additionally, a sample more representative of the general population would not suffer from range restriction in IQ scores. Consequently, the modest effect sizes reported in this investigation (i.e., ≈ |.20| to |.25|) should be regarded as attenuated, in comparison to the results that would likely be obtained from a latent variable analysis with a general community sample. Greater variability in the data would likely lead to larger estimates of internal consistency reliability, as well.

Finally, it is a limitation that the APT was evaluated for concurrent validity with only one other dimension (i.e., psychometric intelligence test scores). It remains to be determined the degree to which scores from the Anagram Persistence Task converge with other putative measures of test-taking persistence. For example, the Mirror
Tracing Task has been used as a measure of test-taking persistence (e.g., Feldman, Dunn, Stemke, Bell, & Greeson, 2014; Quinn, Brandon, & Copeland, 1996). Consequently, the validity of the Anagram Persistence Task, and the construct of persistence more generally, would be established further by the observation of a positive correlation between the APT-Difficult scale and scores from the Mirror Tracing Task. Additionally, it would be useful to determine whether scores from the APT-Difficult subscale relate positively with self-report measures of test-taking effort and persistence (e.g., Arvey et al., 1990; Thelk et al., 2009).

Finally, we note that some researchers include scores from a vocabulary test as a control variable, when using the Anagram Persistence Test, under the assumption that knowledge of vocabulary may impact willingness to persist with an anagram (e.g., Brandon et al., 2003; Quinn et al., 1996). However, based on supplementary correlational analyses, we did not find the Advanced Vocabulary Test scores related more substantially to the APT-Difficult subscale scores, in comparison to the other three intelligence tests. It is possible that it is not verbal ability, per se, that impacts performance on the APT-Difficult subscale. Instead, a confounding factor may be individual differences in self-perception of verbal ability. Future research in this area is encouraged.

**Conclusion**

Previous cautions pertaining to the unnecessary administration of essentially redundant items to represent a dimension have been articulated in the self-report measurement literature (e.g., bloated specific; Cattell & Tsujioka, 1964; Kline, 200). It would appear that there are also behavioral measures for which the administration of several items may be unnecessary. The APT may be one of those measures.
References


DeJoy, D. M. (1985). Information input rate, control over task pacing, and performance
during and after noise exposure. *Journal of Environmental Psychology, 112*(3),
229-242.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M.
(2011). Role of test motivation in intelligence testing. *Proceedings of the
National Academy of Sciences, 108*(19), 7716-7720.

investigation of conscientiousness in the prediction of job performance:
Examining the intercorrelations and the incremental validity of narrow traits.
*Journal of Applied Psychology, 91*, 40–57

267.

training, and goal structure on task persistence. *Journal of Research in
Personality, 26*, 258–272.


Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale:
Pearson, Cronbach, or Spearman-Brown?. *International Journal of Public


Eysenck, H. J. (1988). The concept of “intelligence”: Useful or useless?.
*Intelligence, 12*, 1-16.

Feldman, G., Dunn, E., Stemke, C., Bell, K., & Greeson, J. (2014). Mindfulness and
rumination as predictors of persistence with a distress tolerance task. *Personality and Individual Differences, 56,* 154-158.


cognitive tasks do not produce reliable individual differences. *Behavior Research
Methods*, 1-21.


Hopfenbeck, T. N., & Kjænsli, M. (2016). Students’ test motivation in PISA: the case of

measurement of motivation. *Canadian Journal of Psychology/Revue Canadienne
de Psychologie, 19*(4), 265-279.

Theory of the Relationship Between Individual Differences and Information

failure. In N. Feather (Ed.), Expectations and actions: Expectancy-value models
in psychology (pp. 207- 237). Hillsdale, NJ: Lawrence Erlbaum.


sorting test-64 card version (WCST-64). *Odessa, FL: Psychological Assessment
Resources*.

or a number of independent processes?. *Intelligence, 15*(4), 397-422.

Kubinger, K.D. (2002). Psychology’s challenge when personality questionnaires are
applied for individual assessment. Psychologische Beiträge, 44, 3-9.


Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain


Sommer, M., & Arendasy, M. E. (2016). Does trait test anxiety compromise the measurement fairness of high-stakes scholastic achievement tests?. *Learning and Individual Differences, 50*, 1-10.


Velicer, W.F. (1976). The relation between factor score estimates, image scores, and
principal component scores. *Educational and Psychological Measurement, 36*(1), 149-159.


Table 1

Anagram Persistence Test Inter-Item Correlations and Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>1. M</th>
<th>2. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. APT1</td>
<td>9.58</td>
<td>4.83</td>
</tr>
<tr>
<td>2. APT2</td>
<td>8.10</td>
<td>2.91</td>
</tr>
<tr>
<td>3. APT3*</td>
<td>68.81</td>
<td>35.91</td>
</tr>
<tr>
<td>4. APT4</td>
<td>10.92</td>
<td>6.50</td>
</tr>
<tr>
<td>5. APT5*</td>
<td>61.45</td>
<td>34.95</td>
</tr>
<tr>
<td>6. APT6</td>
<td>11.83</td>
<td>7.93</td>
</tr>
<tr>
<td>7. APT7*</td>
<td>52.55</td>
<td>32.24</td>
</tr>
<tr>
<td>8. APT8*</td>
<td>52.04</td>
<td>32.77</td>
</tr>
<tr>
<td>9. APT9</td>
<td>9.20</td>
<td>4.58</td>
</tr>
<tr>
<td>10. APT10</td>
<td>5.56</td>
<td>1.95</td>
</tr>
<tr>
<td>11. APT11*</td>
<td>33.83</td>
<td>24.11</td>
</tr>
<tr>
<td>12. APT12</td>
<td>8.49</td>
<td>5.35</td>
</tr>
<tr>
<td>13. APT13*</td>
<td>39.12</td>
<td>24.50</td>
</tr>
<tr>
<td>14. APT14*</td>
<td>36.43</td>
<td>26.68</td>
</tr>
</tbody>
</table>

Note. N = 173; APT = Anagram Persistence Test item; correlations ≥ .15 were p < .05; the APT-Easy items were 1, 2, 4, 6, 9, 10, 12; the APT-Difficult items were 3, 5, 7, 8, 11, 13, 14 (highlighted with *).
Table 2

*Standardized Pattern Matrix Loadings and Communalities ($h^2$) Associated with the Anagram Persistence Task Item Factor Analysis*

<table>
<thead>
<tr>
<th>Item</th>
<th>APT-Difficult</th>
<th>APT-Easy</th>
<th>$h^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>APT3</td>
<td>.64</td>
<td>-.17</td>
<td>.42</td>
</tr>
<tr>
<td>APT5</td>
<td>.87</td>
<td>-.20</td>
<td>.77</td>
</tr>
<tr>
<td>APT7</td>
<td>.81</td>
<td>-.05</td>
<td>.65</td>
</tr>
<tr>
<td>APT8</td>
<td>.79</td>
<td>.08</td>
<td>.64</td>
</tr>
<tr>
<td>APT11</td>
<td>.56</td>
<td>.25</td>
<td>.39</td>
</tr>
<tr>
<td>APT13</td>
<td>.53</td>
<td>.16</td>
<td>.32</td>
</tr>
<tr>
<td>APT14</td>
<td>.58</td>
<td>.10</td>
<td>.35</td>
</tr>
<tr>
<td>APT1</td>
<td>-.06</td>
<td>.28</td>
<td>.08</td>
</tr>
<tr>
<td>APT2</td>
<td>-.02</td>
<td>.46</td>
<td>.21</td>
</tr>
<tr>
<td>APT4</td>
<td>.15</td>
<td>.32</td>
<td>.13</td>
</tr>
<tr>
<td>APT6</td>
<td>.18</td>
<td>.26</td>
<td>.11</td>
</tr>
<tr>
<td>APT9</td>
<td>.03</td>
<td>.21</td>
<td>.04</td>
</tr>
<tr>
<td>APT10</td>
<td>-.06</td>
<td>.57</td>
<td>.33</td>
</tr>
<tr>
<td>APT12</td>
<td>.05</td>
<td>.25</td>
<td>.07</td>
</tr>
</tbody>
</table>

*Note.* $N = 173$; principal axis factoring with promax rotation; items $>|.20|$ in bold.
Table 3

Descriptive Statistics and Pearson Correlations between Anagram Persistence Task Scales (Difficult-Unsolved) and General Intelligence Test Performance

<table>
<thead>
<tr>
<th># of Anagrams</th>
<th>$r$</th>
<th>95%CI</th>
<th>$r_c$</th>
<th>$M$</th>
<th>$SD$</th>
<th>Reliability</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.21</td>
<td>.06/.34</td>
<td>.26</td>
<td>68.81</td>
<td>35.91</td>
<td>.42</td>
<td>.24</td>
<td>-1.29</td>
</tr>
<tr>
<td>2</td>
<td>.22</td>
<td>.07/.36</td>
<td>.28</td>
<td>65.13</td>
<td>32.99</td>
<td>.85</td>
<td>.30</td>
<td>-1.15</td>
</tr>
<tr>
<td>3</td>
<td>.19</td>
<td>.05/.34</td>
<td>.24</td>
<td>62.02</td>
<td>30.80</td>
<td>.86</td>
<td>.33</td>
<td>-1.02</td>
</tr>
<tr>
<td>4</td>
<td>.19</td>
<td>.04/.33</td>
<td>.24</td>
<td>59.30</td>
<td>29.10</td>
<td>.88</td>
<td>.40</td>
<td>-.81</td>
</tr>
<tr>
<td>5</td>
<td>.20</td>
<td>.05/.34</td>
<td>.25</td>
<td>57.22</td>
<td>27.82</td>
<td>.91</td>
<td>.40</td>
<td>-.76</td>
</tr>
<tr>
<td>6</td>
<td>.21</td>
<td>.06/.35</td>
<td>.26</td>
<td>54.60</td>
<td>25.56</td>
<td>.89</td>
<td>.39</td>
<td>-.70</td>
</tr>
<tr>
<td>7</td>
<td>.22</td>
<td>.07/.36</td>
<td>.28</td>
<td>52.19</td>
<td>24.65</td>
<td>.86</td>
<td>.47</td>
<td>-.50</td>
</tr>
</tbody>
</table>

*Note.* For the correlations with general intelligence ($r$), $N = 173$; for the internal consistency reliabilities, $Ns$ varied across the seven APT scales, as the number of unsolved anagrams varied across the essentially impossible anagram items (1 $N = 173$; 2 $N = 173$; 3 $N = 157$; 4 $N = 156$; 5 $N = 93$; 6 $N = 86$; 7 $N = 83$); general intelligence was represented by regression-based principal component scores; internal consistency reliability for the general intelligence component scores was estimated via theta ($\Theta = .63$; Armor, 1973); $r_c = correlation disattenuated for imperfect reliability in the general intelligence test scores; reliability for the single-anagram APT-Difficult scale corresponded to the anagram’s communality; reliability for the double-anagram APT-Difficult scale scores was estimated via the Spearman-Brown formula; reliability for the multi-anagram APT-Difficult scale scores was estimated via coefficient alpha.
Table 4

Pearson Correlations between the Anagram Persistence Task Items and General Intelligence

<table>
<thead>
<tr>
<th>Item</th>
<th>r</th>
<th>95%CI</th>
<th>$r_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>- .28</td>
<td>-.41/- .15</td>
<td>-.35</td>
</tr>
<tr>
<td>2</td>
<td>- .21</td>
<td>-.36/- .06</td>
<td>-.26</td>
</tr>
<tr>
<td>3</td>
<td>.21</td>
<td>.05/.34</td>
<td>.26</td>
</tr>
<tr>
<td>4</td>
<td>- .22</td>
<td>-.34/- .10</td>
<td>-.28</td>
</tr>
<tr>
<td>5</td>
<td>.20</td>
<td>-.02/.26</td>
<td>.15</td>
</tr>
<tr>
<td>6</td>
<td>-.20</td>
<td>-.35/- .05</td>
<td>-.25</td>
</tr>
<tr>
<td>7</td>
<td>.05</td>
<td>-.10/.18</td>
<td>.06</td>
</tr>
<tr>
<td>8</td>
<td>.12</td>
<td>-.02/.26</td>
<td>.15</td>
</tr>
<tr>
<td>9</td>
<td>-.08</td>
<td>-.25/.09</td>
<td>-.10</td>
</tr>
<tr>
<td>10</td>
<td>-.22</td>
<td>-.37/- .05</td>
<td>-.28</td>
</tr>
<tr>
<td>11</td>
<td>-.17</td>
<td>-.32/- .02</td>
<td>-.21</td>
</tr>
<tr>
<td>12</td>
<td>-.04</td>
<td>-.19/.11</td>
<td>-.05</td>
</tr>
</tbody>
</table>

Note. Correlations in bold were statistically significant, $p < .05$; $r_c$ = correlations were disattenuated for imperfect reliability in the general intelligence component scores.
Figure 1. Response times (seconds) associated with the Anagram Persistence Task (7 easy anagrams and 7 difficult anagrams) across approximately corresponding presentation order; Panel A = includes all reaction times (solved correctly and incorrectly); Panel B = excludes reaction times from difficult anagrams solved correctly.
Figure 2. Scatter plot depicting the linear/curvilinear associations between Anagram Persistence Task (APT) scores (average response time in seconds) and general intelligence test scores (standardized component scores).