MECHANISMS AND IMPLICATIONS OF RETEST EFFECTS IN RAVEN'S ADVANCED PROGRESSIVE MATRICES

By

Erin R. Neaton

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology – Master of Arts

2023

ABSTRACT

Retest effects are common in higher order cognitive tasks, reflecting the effects of practice. One such task is Raven's Advanced Progressive Matrices (Raven's), the gold standard for tests of fluid intelligence. This study examines two questions concerning retest effects in Raven's: whether the underlying mechanisms include item memory, strategy learning, or both, and whether learning on Raven's affects its validity as a predictive test. We conducted a two-session, remotely administered study in which participants performed either identical Raven's forms in each session, alternate Raven's forms in each session, or a control task in Session 1 and Raven's in Session 2. Raven's form was fully counterbalanced. At the end of Session 2, participants completed tests of fluid intelligence. Results suggest strategy learning, not item memory, is responsible for retest effects. Additionally, correlations between Raven's and the criterion tasks increased between sessions. The experimental results suggest strategy learning may be responsible for this increase, although transient error across sessions may also play a role.

Keywords: Raven's Advanced Progressive Matrices, retest effects, predictive validity

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Zach Hambrick, Erik Altmann, and Kimberly Fenn for serving on my committee and providing invaluable guidance and feedback throughout this research.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

A major focus of psychometric research on human intelligence is to identify statistical factors that account for individual differences in cognitive ability (6). This research has established that scores on diverse tests of cognitive ability tend to correlate positively with each other (18; 13). This "positive manifold" implies the existence of a general factor of intelligence—psychometric *g*. However, this work has further established the existence of cognitive ability factors that are more specific than *g* but still relatively broad (7; 12). Cattell (1943) defined fluid intelligence (Gf) as the ability to solve novel problems and crystallized intelligence (Gc) as knowledge or skill acquired through experience.

Raven's Advanced Progressive Matrices is the gold standard test of Gf. It has been used to test participants in thousands of research studies, to identify qualified employees, and to select gifted children (3; 1). The goal in each item on Raven's is to determine the missing element in a pattern which is presented in a matrix, and each item is designed to be harder than the previous item (15). Except for the test instructions, which can be translated to nearly any language, Raven's is entirely nonverbal. Thus, at least in theory, people who speak different languages can be compared on Raven's scores. Raven's can also be used for a wide range of ages because there is no minimum reading level requirement.

As with many other tests of cognitive ability, test takers tend to improve in their performance on Raven's across multiple administrations (14; 5; 4). It is unclear, though, what drives these *retest effects*. Two potential mechanisms are *item memory* and *strategy learning*. Item memory occurs when a participant remembers their answers to items from a previous administration of a test. Raven's is a timed test, so remembering answers would enable participants to answer items more

quickly and thus to attempt more items. Since no feedback is provided in Raven's, there is no reason to expect that participants would have memory for answer accuracy. As a result, if item memory is responsible for retest effects, we would expect faster response times, rather than improved accuracy on repeated items, to be the main source of improvement, as participants could reach more items. To our knowledge, no study to date has investigated the role of item memory as a potential mechanism underlying retest effects in Raven's.

By contrast, strategy learning occurs in a test when participants develop a particular, systematic approach to responding to items. In a previous study, participants were trained to identify patterns across rows and then patterns across columns, and then to use this information to determine which answer option followed both patterns (8). Other work has demonstrated that a variety of strategies can be learned, some without the need for explicit training (20; 19; 10). In strategy learning, a strategy is learned as people work through the items. Once it is learned, it can be applied to the subsequent items and on subsequent administrations of the task. This increases the speed at which people can work through the items, as well as the accuracy on the items.

Based on this previous research, we expect that strategy learning contributes to retest effects, but it is important to determine whether item memory also contributes. This is of interest to inform our theoretical understanding of retest effects, and it will also inform how we can design tasks and studies to mitigate retest effects in practice.

A related issue that has received little, if any, empirical attention is how retest effects impact the *predictive validity* of Raven's. Predictive validity refers to the strength of the correlation between a predictor variable and a criterion variable. Retest effects could change the rank ordering of participants in Raven's from the first administration to the second, and thereby change the test's predictive validity. For example, if item memory explains retest effects in Raven's, then

participants higher in episodic memory ability would be expected to rank higher among participants on the second administration of Raven's than on the first. In this situation, Raven's scores would also be expected to correlate less with other measures of Gf on the second administration than on the first. It is especially important to understand how retest effects influence the predictive validity of Raven's, given that the test is used as a predictor of job performance and academic potential, and a measure of cognitive ability which factor into real-world opportunities for individuals (3; 1).

CHAPTER 2

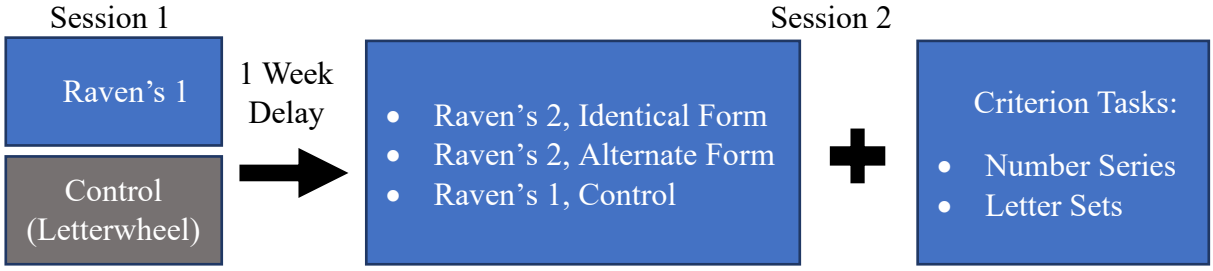
PRESENT STUDY

The major question of this study was whether retest effects in Raven's reflect item memory or strategy learning, or both mechanisms. To answer this question, we conducted an online study utilizing lab.js to compare retest effects in groups of participants who completed either identical or alternate Raven's forms across two sessions (11). Participants were randomly assigned to one of three conditions: Identical Form, Alternate Form, or Control. The Identical Form and Alternate Form groups both completed Raven's in Session 1, whereas the Control Group completed a complex procedural task (the Letterwheel task) in Session 1 (2).

In Session 2, which occurred after a delay of approximately one week, all three groups completed Raven's and two criterion tasks. The groups differed in which Raven's form they completed at Session 2. The Identical Form Group completed the same form of Raven's in Session 1 and Session 2 (i.e., the Odd Form in both sessions or the Even Form in both sessions). The Alternate Form group completed different forms of Raven's in Session 1 and Session 2 (i.e., the Odd Form in Session 1 and the Even Form in Session 2, or vice-versa). Participants in the Control Group did not complete Raven's at Session 1, so they saw the form they completed for the first time at Session 2. Forms were counterbalanced between participants in the Identical Form and Control Groups. Form order was counterbalanced across sessions in the Alternate Form Group.

Figure 1

Illustration of the Study Design



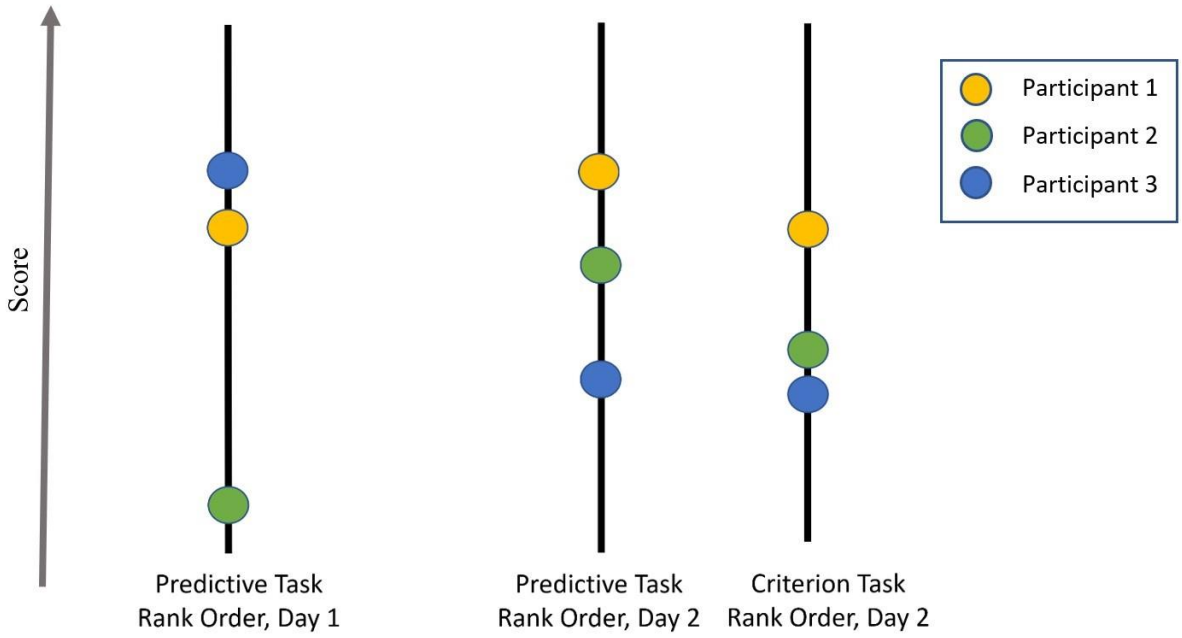
Note. Layout for the study design for the mechanisms behind retest effects and the predictive validity of Raven's.

A significant retest effect in the Identical Form group would indicate that item memory, strategy learning, or both occurred, whereas a significant retest effect in the Alternate Forms group would indicate that strategy learning occurred. Finally, a significant retest effect in both groups, but a larger effect in the Identical Form group than in the Alternate Form group, would indicate that both mechanisms contribute to retest effects in Raven's.

The second question was whether retest effects impact the predictive validity of Raven's. To answer this question, we compared correlations of Raven's at each administration with scores on two additional tests of Gf administered in Session 2, which served as criterion variables. These criterion variables were combined into a composite. We expected that the criterion task composite would correlate more strongly with Session 2 Raven's scores than with Session 1 Raven's scores due to transient error. Transient error is defined as the "longitudinal variations in responses to measures that are produced by random variations in respondents' psychological states across time" (17). It is the error resulting from larger intraindividual variability for tasks completed on separate days rather than the smaller intraindividual variability for tasks completed on the same day. Participants are more self-similar on one day during a particular assessment period than across multiple days and assessment periods. This means they likely perform more similarly on tasks they complete in one session than in tasks they complete across different sessions.

Figure 2

Illustration of Transient Error's Impact on Rank Order



The critical question was whether changes in correlations were present above and beyond changes that can be attributed to transient error, and whether those changes were in the direction of stronger or weaker correlations. The differences in correlations between Raven's and the criterion task composite were analyzed both within and across groups. The Identical and Alternate Form Groups were combined for these analyses into an Experimental Group. A significant difference between the Session 1 Raven's and the criterion task composite correlation and Session 2 Raven's and the criterion task composite correlation for the combined Experimental Group indicates the difference can be attributed to either transient error or mechanisms like item memory or strategy learning. If the correlation decreases, it means the predictive ability of Raven's decreases with subsequent administrations. If the correlation increases, it means that either the predictive ability of Raven's increases with subsequent administrations or reduced transient error across sessions significantly increases the correlation.

To tease apart these two possibilities, two comparisons were conducted. The first compares the correlations of Raven's 1 and criterion tasks between the Control Group and combined Experimental Group. Since these Raven's administrations are the first for each group, but are completed on different days, a significant difference would provide evidence for the presence of transient error which results in a higher correlation for the tasks completed in the same session. The second comparison analyzes the correlations of Session 2 Raven's and criterion tasks completed between the Control Group and combined Experimental Group. These tasks were all completed on the same day, so a significant difference here indicates a difference that can be attributed to something other than transient error, such as participants in the Experimental Group having previous experience with Raven's. If the results of these two comparisons align, it provides evidence either for or against the presence of transient error. Otherwise, further research is necessary to clarify the results.

CHAPTER 3

METHOD

3.0.1 Participants

The participants in this study were undergraduates at Michigan State University (MSU), recruited through the MSU Department of Psychology subject pool. The average age of participants was 20.5 years (*SD* = 1.6). Participants received class credit for their participation. A power analysis indicated that group sample sizes of 208 participants would provide sufficient statistical power (99%) to detect moderate effect sizes (*d* = .30). A total of 564 participants completed the study. We removed 9 participants who had completed Raven's prior to this study, leaving N = 555.

3.0.2 Materials

Two Raven's forms were used in this study, the Even Form and the Odd Form. In each form, an item consists of a matrix of nine pieces of a pattern, with the bottom right piece missing. The task is to choose, from eight answer options, the piece that best completes the patterns in the rows and columns. There are 18 items which are designed to get progressively harder. Participants have 10 minutes to complete as many items as possible.

During Session 2, all participants completed two additional tests of Gf. In Number Series, an item consists of a series of five numbers that follow a pattern. The goal is to determine which number out of the five answer options comes next in the pattern. There are 15 items and participants have four and a half minutes to complete as many of these items as possible. In Letter Sets, an item consists of four sets of letters which all follow a pattern and one set which does not follow the pattern. The participant's task is to identify the set of letters from the five options that does not follow the pattern. There are 20 items and participants have five minutes to complete as many items as possible.

3.0.3 Procedure

Participants completed the study online in two sessions. After clicking a link from the MSU subject pool website, participants were taken to a landing page where they were assigned to the Identical Form Group, Alternate Form Group, or Control Group, and then sent to the website where that group's version of the study was hosted. They started by viewing a consent form which they could download. The consent form listed the researcher's contact information in case participants had any issues, questions, or concerns. After viewing the consent form, participants began the study. At the end of Session 1, participants were linked back to the MSU subject pool website and granted half of their credit.

One week later, participants were able to access a link through the MSU subject pool website to complete Session 2 of the study. The link was available for five days after it became available. This link took them to the Session 2 landing page which matched their ID with their subject group and sent them to the website where that group's version of the study was hosted. After completing the tasks in this session, participants viewed a debriefing form, which they could download if they wished. They were then directed back to the MSU subject pool website and granted the second half of their credit.

CHAPTER 4

RESULTS

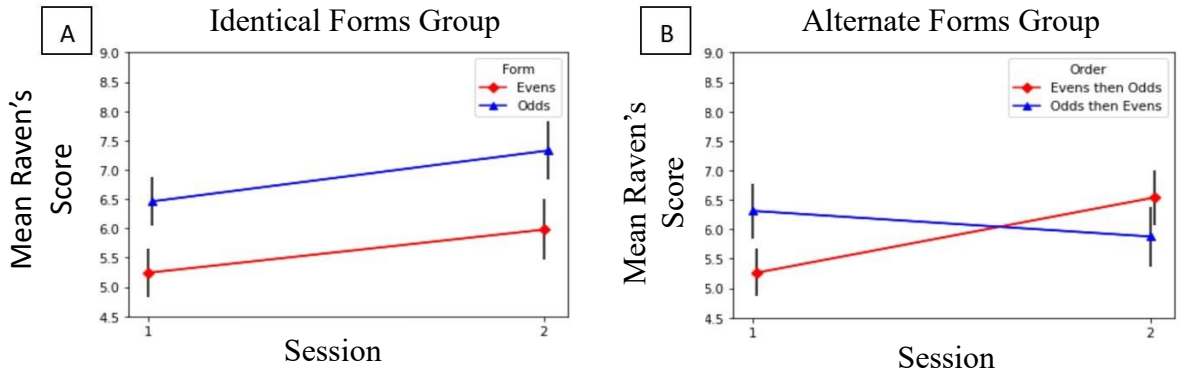
4.0.1 Mechanisms of Retest Effects in Raven's

The first set of analyses investigated the underlying process or processes responsible for any score improvements. If Raven's scores improved due to item memory, we would expect performance to improve in the Identical Form Group. If they improved because of strategy learning, we expect performance to improve in both the Identical Form Group and the Alternate Form Group.

We conducted a 2 (Session: First, Second) x 2 (Group: Identical, Alternate) x 2 (Form: Odd, Even) ANOVA. There was a main effect of Session, $F(1, 382) = 4.83$, $p = .0282$, $\eta^2 = .003$. Mean Raven's score was higher on average in Session 2 than in Session 1. There was also a main effect of Form, $F(1, 382) = 17.60$, $p < .01$, $\eta^2 = .0101$. Mean Raven's score was higher for the Odd form than for the Even form. Finally, there was no Session x Group interaction: $F(1, 382) = 1.27$, $p = .260$, $\eta^2 = .001$. This suggests that the amount of learning that happened in each group, identical or alternate forms, did not significantly differ. This points toward strategy learning because both groups improved, and there was no difference between groups in how much they improved.

Figure 3

Retest Effects Results



Note. Results from Session 1 to Session 2 for the A) Identical Form Group and B) Alternate Form Group. Error bars represent 95% confidence intervals.

4.0.2 Predictive Validity of Raven's

The predictive validity of Raven's in the group that took the assessment twice was analyzed within participants. This was done to test for changes in predictive validity. To control for transient error which may be present due to the criterion tasks being tested at Session 2, two additional comparisons were conducted. The predictive validity of Raven's at the second administration was compared between the group that saw Raven's once (the Control Group) and the second administration of the groups that saw Raven's twice (the Experimental Group). The predictive validity of Raven's 1 and the criterion tasks was also compared between groups.

We calculated the correlations between Raven's scores for each session and the criterion task composite across all participants in the two groups who saw Raven's twice. We then used an online calculator from Lee & Preacher (2013) to calculate a test statistic for the difference between the two dependent correlations. This difference was significant, $r_{\text{Raven's 1, Criterion}} = 0.47$, $r_{\text{Raven's 2, Criterion}} = 0.56$, $z = -2.40$, $p < 0.01$. The correlation between Raven's and the criterion tasks was Significantly stronger at Session 2 than at Session 1.

We calculated the correlations between Session 2 Raven's scores and the criterion task composite for the group that saw Raven's once and compared it with the correlation calculated above for the Experimental Group's Session 2 Raven's scores and the criterion task composite. We followed the same procedure as above, but this time for a test between independent samples (Preacher, 2002). This comparison was conducted to determine between groups differences in correlations for Session 2 Raven's and the criterion tasks between those who take Raven's for the first time at Session 2 and those who take Raven's for the second time at Session 2. This controls for transient error. If there is a significant difference, this would imply the performance at Session 2 for those who have seen Raven's twice improved from Session 1 above and beyond the increase from transient error alone. The difference was not significant, $r_{\text{Raven's Control, Criterion}} = 0.50$, $r_{\text{Raven's Experimental S2, Criterion}} = 0.56$, $z = 1.11$, $p = 0.13$. This result does not provide any evidence that strategy learning from Session 1 to 2 contributed to a change in predictive validity of Raven's.

We also compared the independent correlations between Raven's scores and the criterion task composite between the Control Group's Raven's and the combined Experimental Group's Session 1 Raven's. This comparison aimed to test whether transient error is greater when the criterion tasks and Raven's 1 are completed on different days than when completed on the same day. If it is significant, it would show that transient error between Session 1 and 2 significantly impacts correlations. This difference was not significant $r_{\text{Raven's Control, Criterion}} = 0.50$, $r_{\text{Raven's Experimental S1, Criterion}} = 0.47$, $z = -0.4801$, $p = 0.3153$. This result does not provide evidence that transient error significantly affects the correlations.

The results from these analyses are ambiguous. It is unclear whether the transient error is responsible for a significant amount of the increase in the strength of the correlation between Raven's at Session 1 and 2 and the criterion task composite for the group that saw Raven's twice.

There is a significant increase in the correlation, but it is unclear whether the increase is due to strategy learning, transient error, or a combination of the two. We plan to clarify this ambiguity with a follow-up study described below.

CHAPTER 5

DISCUSSION

Raven's Advanced Progressive Matrices is the gold standard test of fluid intelligence. It is used in research, academic, and industry settings and scores can impact people's opportunities in these settings. Retest effects are known to occur on multiple administrations of Raven's, but little work has analyzed the underlying mechanisms. Here, we present a two-session online study to analyze whether item memory, strategy learning, or both are the mechanisms underlying retest effects and how the predictive validity of the test changes with multiple uses.

5.0.1 Mechanisms of Retest Effects in Raven's

The first major finding of this study is that the learning that happens on Raven's across administrations seems to be from strategy learning, rather than item memory. The evidence we provide is that regardless of whether participants saw the identical form with the same items, or a new form with new items, they improved from Session 1 to Session 2. Additionally, the amount of improvement did not differ between groups and the effect size for this interaction was near zero. This provides evidence that seeing the same items both times does not give participants an advantage, but instead, participants in both groups could perform equally well, likely due to strategy learning.

This result is consistent with results from a study by Gonthier and Thomassin (2015) who found that when participants use strategies on Raven's, the change in strategy use fully mediates the change in correlation between the performance on working memory tasks and Raven's. This mediation reveals that the use of strategic behavior is part of what drives the correlation and provides further evidence that retest effects are due to strategy learning on tasks of Gf.

These results increase scientific understanding of retest effects in cognitive ability testing. They are also relevant to studies of aging. Practice effects are one possible explanation for why results concerning age-related cognitive decline are different in longitudinal and cross-sectional studies (16). Retest effects are a potential explanation for why longitudinal studies seem to indicate cognitive aging occurs later than cross-sectional studies indicate. Research examining the mechanisms of retest effects will provide further clarification on this issue. If strategy learning is responsible for the retest effects, as shown in this study, and if the strategies that are learned cannot be applied to tasks other than Raven's, it indicates the repeated use of Raven's results in inflated scores. Without knowing the reason for the score increase, it would seem people score as high or better in ability on their subsequent administrations, as compared to their first. This result would point to an increase or a lack of decrease in cognitive ability. If we know that strategy learning is responsible for these score increases, as indicated by the evidence from this study, we know the score increase is due to using a strategy specific to this task, rather than an increase in ability.

The results of this study also provide a cautionary note that there is a difference in performance between forms—the mean score on the odd form was significantly higher than the mean score on the even form. This finding is important, especially for studies that involve a design where researchers may want to compare scores longitudinally. Since the forms are not interchangeable, researchers should not compare scores from the two forms at separate times to track changes over time.

5.0.2 Predictive Validity of Raven's

The second major result of this study is that there is not a clear picture of whether the predictive validity of Raven's changes with repeated administrations of Raven's. It is unclear whether the increase in the strength of the correlation between Raven's at Session 1 and 2 and the criterion task

composite is due to strategy learning, transient error, or a combination of the two. In a subsequent study, we will replace the Control Group with a group who completes Raven's twice in a single session, along with the criterion tasks. This will allow us to test how much of the increase in correlation is due to strategy learning. The difference between this group and the group who completes Raven's across two different sessions will elucidate how much of the increase is due to transient error.

Regardless of whether this outcome is due to strategy learning or transient error, the results suggest Raven's can be used as a predictive measure multiple times. The results here show an increase in the predictive validity of Raven's, rather than a decrease. Using the task multiple times may actually result in a better prediction of criterion tasks, depending on how much of the increase transient error accounts for and on the specific criterion task used. This result has major implications for research, school standardized testing, and job ability assessments. Had the correlation decreased across sessions, we would only be able to use Raven's Matrices one time to predict a criterion accurately. If the results here hold when tested with other criterion tasks and varied timelines, they reveal that we can use participants from subject pools in which the same participants take part in many studies. Additionally, the results illustrate we can continue to use Raven's on job assessments and children's aptitude tests.

Previous research has shown that when using working memory tasks to predict Raven's, strategy training interventions for Raven's decrease the correlation between the performance on the two types of tasks (working memory and GF tasks) (19). This finding is different than what we would expect given our results here, but it may be a result of specific tasks used, rather than conflicting results. In some cases, the strategies can account for the majority of score gains that are present in retest effects and their correlations with criterion tasks (9; 10). The strategy learning that

enables improvement on one task seems to only transfer to tasks that are very similar to the original task. Otherwise, the correlations between that task and criterion tasks tend to decrease. In the current study, the correlation likely increased because Raven's was very similar to the two criterion tasks—all three are reasoning tasks which are used to evaluate Gf.

CHAPTER 6

CONCLUSION

The results presented here provide evidence that retest effects in Raven's are a result of strategy learning, and not item memory. This finding helps improve our understanding of retest effects and helps clarify the implications for longitudinal studies. Additionally, it remains unclear how these retest effects impact the predictive validity of Raven's. The correlations between Raven's and the criterion composite increase significantly across time, but a follow-up study is necessary to determine whether this increase is a result of transient error or strategy learning.

The findings from this study have wide-reaching implications. In research studies, the results here indicate that we can use Raven's multiple times to predict at least certain criterion task performance, but we should not use Raven's score changes to track changes in cognitive ability over time without controlling for practice effects. For academic and industry settings, Raven's seems to predict criterion tasks well with one use. Further studies should analyze whether this holds true for criterion tasks in other domains with multiple administrations.

BIBLIOGRAPHY

[1] Raven's progressive matrices: Job assessment test. https://oya-aptitude-test.com/ ravens-progressive-matrices/.

[2] Altmann, E. M. and Hambrick, D. Z. (2022). Task independence of placekeeping as a cognitive control construct: Evidence from individual differences and experimental effects. *Cognition*, 229:105229.

[3] Balboni, G., Naglieri, J. A., and Cubelli, R. (2010). Concurrent and predictive validity of the raven progressive matrices and the naglieri nonverbal ability test. *Journal of Psychoeducational Assessment*, 28(3):222–235.

[4] Belacchi, C., Carretti, B., and Cornoldi, C. (2010). The role of working memory and updating in coloured raven matrices performance in typically developing children. *European Journal of Cognitive Psychology*, 22(7):1010–1020.

[5] Bors, D. A. and Vigneau, F. (2003). The effect of practice on raven's advanced progressive matrices. *Learning and Individual Differences*, 13(4):291–312.

[6] Carroll, J. B. et al. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge University Press.

[7] Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological bulletin*, 40(3):153.

[8] Denney, N. W. and Heidrich, S. M. (1990). Training effects on raven's progressive matrices in young, middle-aged, and elderly adults. *Psychology and aging*, 5(1):144.

[9] Gonthier, C. and Thomassin, N. (2015). Strategy use fully mediates the relationship between working memory capacity and performance on raven's matrices. *Journal of Experimental Psychology: General*, 144(5):916.

[10] Hayes, T. R., Petrov, A. A., and Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, 48:1–14.

[11] Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., and Hilbig, B. E. (2021). lab. js: A free, open, online study builder. *Behavior Research Methods*, pages 1–18.

[12] Horn, J. L. and Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, 57(5):253.

[13] Jensen, A. R. (1998). The g factor. *Westport, CT: Prager*.

[14] Nkaya, H. N., Huteau, M., and Bonnet, J.-P. (1994). Retest effect on cognitive performance on the raven-38 matrices in france and in the congo. *Perceptual and Motor Skills*, 78(2):503–510.

[15] Raven, J. C. and Court, J. (1938). Raven's progressive matrices. *Western Psychological Services: Los Angeles, CA*.

[16] Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*, 24(5):563.

[17] Schmidt, F. L., Le, H., and Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological methods*, 8(2):206.

[18] Spearman, C. (1961). " general intelligence" objectively determined and measured.

[19] te Nijenhuis, J., van Vianen, A. E., and van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, 35(3):283–300.

[20] Turley-Ames, K. J. and Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of memory and language*, 49(4):446–468.