5

Rasch Analysis of the Purdue Nonverbal Test: Any Use for Ancient Tools in the Modern Era?

Siniša Lakić
Department of Psychology, Faculty of Philosophy, University of Banja Luka
sinisa.lakic@ff.unibl.org

Biljana Mirković
Department of Psychology, Faculty of Philosophy, University of Banja Luka

Lana Vujaković
Department of Psychology, Faculty of Philosophy, University of Banja Luka

Abstract

Among the instruments used by psychologists in the Western Balkans, the Purdue Nonverbal Test (PNT) of visual perceptual abilities stands out as an iconic tool. Despite its venerable age of over 60 years and the clearly outdated norms questioning its validity, psychologists in Bosnia and Herzegovina continue to employ it, especially in driving ability assessments. With this in mind, we aimed to evaluate the test's functionality by establishing actual norms for new drivers via Rasch analysis. The study involved 721 participants (52.1% male), all of whom were final-year high school students. In addition to the PNT, participants completed the Short Sensation Seeking Scale (from SUPPS-P) and a specially designed Scale of Attitudes Towards Risky Driving Behaviours for this study. Analysis of the items' unidimensionality, local independence, and item fit exposed issues with three items, which were subsequently removed from further analysis. The corrected version showed no gender bias and no indication of items with significant differential functioning. Criterion validity analysis showed that participants with a driving license significantly outperformed those without one, together with expected differences appearing across high schools attended. We observed insignificant correlations with risky driving behaviours and sensation seeking, which aligns with theoretical expectations. In addition, the information value of the PNT peaked at lower levels of ability which is vital for assessing deficient driver candidates. We discuss the benefits of employing Rasch analysis for crafting similar selection instruments and adapting time-honoured tools in the contemporary era.

Keywords: *Rasch analysis, cognitive abilities, driving, psychometrics, Purdue Nonverbal Test*[1]

[1] The study was funded by the Traffic Safety Agency of the Republic of Srpska and represents a substantially updated English-language version of an internal technical report prepared for the Agency.

Introduction

After instances of near-miss traffic incidents, drivers frequently make snap judgments about other drivers' cognitive deficiencies. While such assessments may reflect bias, since non-cognitive personality attributes might better explain risky driving behaviours (see e.g., Bowen et al., 2020), there is substantial empirical evidence supporting lay theories of the association between cognitive abilities and hazardous traffic behaviours. There are early studies indicating a direct link between cognitive functioning and accident rates (e.g., Arthur et al., 1991; McKnight & McKnight, 1999), studies that show associations with driver training success and job performance (e.g., Bertua et al., 2005; Salgado et al., 2003), and studies conducted in specialised driving simulators linking cognitive abilities and risky behaviours in virtual traffic scenarios (e.g., Casutt et al., 2014; Kaber et al., 2016; and recently Collins, 2023 for train drivers). Given that cognitive decline is a natural consequence of the ageing process and of a number of medical conditions, including schizophrenia, dementia, and multiple sclerosis, one should not be surprised that a significant portion of this research has focused on determining the relationship between driving behaviour and cognitive functioning within these subpopulations (e.g., Mathias & Lucas, 2009; Reger et al., 2004; Schultheis et al., 2010).

The convergence of commonsense knowledge and scientific evidence has prompted many countries to introduce the evaluation of cognitive abilities when obtaining or renewing a driving license. However, legal regulations vary significantly among countries. In the majority of them psychological assessment which involves cognitive testing is mandated only for specific candidates: those with certain medical conditions, history of substance use, history of prior traffic violations, professional drivers, and those who have reached a certain age. Only a small number of countries mandate a psychological assessment for all cases of issuance and/or renewal of driving licenses (e.g., Spain, Brazil). Bosnia and Herzegovina is among them.

In particular, the Regulation on Health Requirements for Motor Vehicle Drivers in Bosnia and Herzegovina (Pravilnik, 2007) distinguishes among three categories of drivers, and accordingly specifies three different cognitive ability cut-offs. For non-professional drivers—who constitute the largest number of candidates—the disqualifying conditions are defined as "forms of intellectual insufficiency below borderline values, regardless of cause." For professional drivers whose primary occupation is operating motor vehicles, the disqualifying conditions are defined as "all forms of reduced intellectual ability classified as borderline or below, regardless of aetiology." Finally, for Category D drivers (commercial passenger vehicle operators), intellectual abilities "must not be below average." The consistency of such a classification is open to debate since one needs to define the cut-offs operationally. However, from the perspective of a psychometrician and applied practitioner, the choice of assessment instruments could be of even greater importance.

According to our informal inquiry of practising psychologists in the Republic of Srpska (a self-governing entity within Bosnia and Herzegovina), the Purdue Nonverbal Test (PNT; originally called Purdue Non-Language Test, Tiffin et al., 1958) appears to be the first choice among cognitive ability assessment tools. The PNT is a non-verbal test designed to evaluate visual-perceptual cognitive abilities while purportedly being "culture-free". The test comprises two parallel forms (A and B), each consisting of 48 items. Notably, our literature search revealed very few studies examining the instrument's psychometric properties, yet it seems to have gained considerable popularity in applied psychology in the former Yugoslavia already since the 1960s (according to Dautović & Borčić-Konjarek, 1998) and it has continued to be used in its successor states. For example, in one of two found studies from the region, Piri (2007) reports that at least until 2006, psychologists in Croatia—Bosnia and Herzegovina's neighbouring country—systematically employed the PNT to test driver candidates.

This is despite the fact that almost a decade earlier, Dautović and Borčić-Konjarek (1998), questioned the appropriateness of the then-used norms based on a study of 500 driving licence applicants. Their primary criticism centred on the asymmetric distribution of scores across all age groups, resulting from the test's easiness

for all except the oldest cohort (age 54 and above). These authors argued for revisions focusing on increasing the number of moderately difficult and difficult items to "enhance discrimination at higher score ranges." As will be seen in the remainder of this paper, we will demonstrate why we believe that a negatively skewed distribution is actually desirable for the intended purpose of the test in the context of drivers' evaluation. Furthermore, we will discuss how, rather than increasing the number of items, it would be more beneficial to make PNT-like tests more efficient through the use of modern technologies that were not available at the time.

The second important study on the PNT conducted in Croatia was carried out by Piri (2007). The author collected archival records of 413 drivers under 30 years of age and 237 drivers aged 65 and above. Rather than focusing on internal psychometric properties, this study examined the criterion validity of the scores registered at psychological assessment in 2002. Piri reported that, in the younger age group, PNT scores negatively correlated with the number of attempts required to pass the theoretical driving examination. However, no significant relationships were found with practical driving test performance, nor were there associations in either age group between PNT scores and legally documented traffic violations or accidents during the approximately four-year follow-up period. While these findings might appear to challenge the functionality of the PNT, the author convincingly argues that the reasons for the obtained null effect could have been the relative rarity of documented violations, the presence of a large number of external variables influencing both the passing of the practical test and traffic incidents, and the sample selection bias, as the study excluded individuals denied license due to intellectual dysfunction.

All in all, despite understandable reservations about examining the functioning of a test that is over 60 years old, we believe there are at least two reasons why this endeavour is valuable. Firstly, while its persistent use in practice likely stems from the limited availability of superior commercial alternatives in Bosnia and Herzegovina, we posit that practitioners would not continue to use the same instrument for decades unless it maintained some utility. Whether this is the case needs to be determined empirically.

Secondly, the aforementioned criticism directed at the PNT regarding the non-normal distribution of scores stems from reliance on the postulates of formulaic application of classical test theory. For this reason, this paper aims to present to interested readers the application of Rasch modelling, which is more appropriate for ability tests and adheres more consistently to the fundamental principles of developing psychological instruments—that one must always be aware of the purpose of testing, and that generic solutions are not always optimal.

With the above in mind, we investigated the psychometric properties of the PNT by exploring (1) its internal structure (unidimensionality, local independence, and internal consistency) and (2) its relationships with theoretically relevant external variables. As for criterion validity, we tested several hypotheses. To support its utility in drivers' testing, the PNT should show the absence of gender bias and higher performance among those who have already become drivers. As a cognitive ability measure, the PNT should discriminate among students from schools with varying academic demands. Finally, based on theoretical considerations and recent empirical findings (Anglim et al., 2022; Stanek & Ones, 2023), one should observe minimal associations with relevant non-cognitive attributes, specifically sensation seeking and its derivative, attitudes towards risky driving.

Method

Participants and Procedure

Given that the study was primarily aimed at testing the adequacy of the PNT for potential drivers, we selected high school seniors as the operational population. Almost all high school seniors in Bosnia and Herzegovina are 18 years old, meeting the age requirement for taking the driving test. Not only do they rep-

resent typical candidates for a Category B driving license, but many of them have also had the opportunity to obtain an A1 category license starting at the age of 16. Indeed, out of the total sample of 721 valid responses, nearly half (340, 49.6%) reported already holding a driving license (ranging from categories A1 to CE, with the majority holding category B).

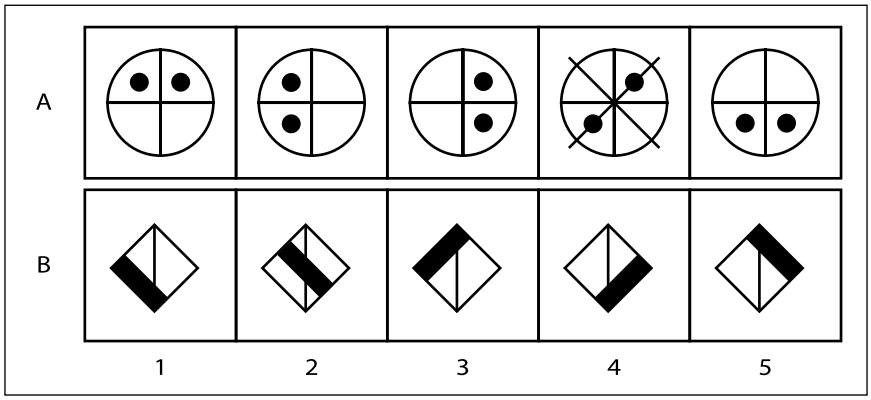
To better reflect the variability in cognitive abilities within the population, we included six high schools in the city of Banja Luka (the largest city in the Republic of Srpska), representing a range of academic achievement among students. Specifically, we included the Gymnasium (*n* = 231), Electrotechnical School (*n* = 94), and Music School (*n* = 13), which enrol students with higher academic performance in primary school, as well as the Polytechnic (n = 94), Technical (*n* = 172), and Agricultural School (*n* = 117), which tend to enrol students with lower educational achievement. This selection resulted in a slightly higher number of male respondents (376, 52.1%) compared to females (300, 41.5%), with 46 participants (6.4%) choosing not to disclose their gender. Although the sample was not perfectly balanced, this gender distribution more closely reflected the gender distribution of drivers in the Republic of Srpska, which is approximately 2:1 male to female (BIHAMK, 2018).

Data collection took place during 2021, after approvals were obtained from the Ministry of Education and Culture of the Republic of Srpska and from each individual school. A member of the research team, in collaboration with the school psychologist or pedagogue, administered the battery in a group-setting following a precisely defined protocol during a single 45-minute school class.

Measures

As should be clear from above, the primary focus of this study was the Purdue Nonverbal Test, specifically its Form A. The test consists of 48 nonverbal items. Each item presents five drawings composed of geometric shapes, where one drawing deviates from others by not following the pattern applied to the remaining combinations. Sample items, used for practice before independent work, are shown in Figure 1.

Figure 1
Initial practice items from the PNL test



The Short Sensation Seeking Scale was taken from the SUPPS-P instrument (see Cyders et al., 2014). SUPPS-P is a short version of a longer impulsivity assessment tool (UPPS-P; Lynam et al., 2006), which, in addition to sensation seeking, includes four other subscales: positive urgency, negative urgency, lack of perseverance, and lack of premeditation. The shortened versions comprise four items, to which participants respond using a 4-point Likert scale (ranging from *not at all* to *completely*). Example items for the sensation-seeking scale include: "*I quite enjoy taking risks*" and "*I welcome new and exciting experiences and sensations, even if they are a little frightening and unconventional*" Given the small number of items, the internal consistency coefficients were expectedly low, but reasonably satisfactory (α = .67). Other studies have shown adequate

psychometric characteristics of the scale, including its test-retest reliability (Dugre et al., 2019).

The Scale of Attitudes Towards Risky Driving Behaviours was developed for the purposes of this project, and the final version consisted of 20 items. The participants were asked to rate, on a 4-point scale (ranging from *never* to *often*), to what extent they believe certain, generally risky behaviours are justified for drivers in traffic. The statements were phrased in a way that tried to minimise socially desirable responding through positive framing. Example items include: "*Having a phone conversation while the vehicle is in motion, holding the phone in your hand, if the person is confident in their abilities*" and "*Driving aggressively with sudden accelerations and braking to demonstrate to less skilled drivers how to drive efficiently and alleviate traffic congestion*" The first unrotated factor accounted for 31% of the variance in the scores, with all items significantly loaded on it ($\lambda > .32$). The internal consistency of the summation score was high ($\alpha = .90$). The correlation with the Short Sensation Seeking Scale was moderate ($r = .31$, $BF_{10} > 1000$, $p < .001$), suggesting that they measure different constructs.

Data Analysis

After double data entry, the quality of the data was checked by analysing the theoretically possible range of values and using techniques to check for response patterns' plausibility (i.e., improbable runs of same responses, psychometric synonyms and antonyms scores). Clear indications of demotivated responding were observed for 14 participants, who were excluded from further analysis, resulting in a final sample of 721 participants. Given the advantages of the Rasch model in psychometric analysis of abilities tests, the techniques within this approach were the focus of the analysis. Specifically, the unidimensionality of the PNT was primarily tested using the Martin-Löf test, Ponocny's T_{1m} test, calculating the percentage of variance explained by the Rasch model, and conducting a principal component analysis on the residuals (see, e.g., Smith, 1996). Local independence was tested using Yen's Q3 measure, and item difficulty and discrimination were complemented by fit analyses of items and respondents using the Rasch model. Criterion validity was assessed by examining the relationship with variables such as gender, school type, possession of a driving license, sensation-seeking scales, and attitudes towards risky driving behaviours. Additionally, differential item functioning was analysed based on gender, and the informativeness of the test at different levels of ability was examined. Finally, score norming was performed, and optimisation was conducted for more efficient ability estimation. All analyses were conducted in R environment using the following packages: ltm (Rizopoulos, 2006), eRm (Mair & Hatzinger, 2007), perfit (Tendeiro & Meijer, 2014), sirt (Robitzsch, 2020), difR (Magis et al., 2010), TAM (Robitzsch et al., 2020), WrightMap (Torres Irribarra & Freund, 2014), eatATA (Becker et al., 2021), mirt (Chalmers, 2012), and ufs (Peters, 2021).

Results

After it had been established that each item exhibited at least minimal variability (with item difficulty ranging from .28 to .99), we performed analyses based on a factor model. Interestingly, none of the four frequently recommended methods for determining the number of common factors in exploratory factor analysis (Horn's parallel analysis, Horn, 1965; MAP, Velicer, 1976; Hull's method, Lorenzo-Seva et al., 2011; and Empirical Kaiser Criterion, Braeken & van Assen, 2017) suggested a unidimensional solution. The number of proposed factors ranged from two to seven, depending on the method and the type of correlation coefficient used for analysing the correlation matrix (phi, tetrachoric, or gamma).

However, the first unrotated factor was dominant, accounting for 34% of the total variability and saturating 46 out of 48 items with a value greater than 0.32 (which is frequently considered the recommended lower threshold, Costello & Osborne, 2005). Further analysis revealed a strong effect of item dif-

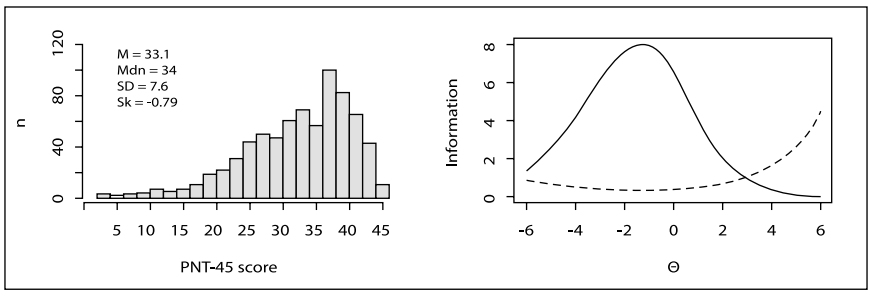
ficulty on the factor solution. Specifically, it was observed that the first two rotated components were relatively highly correlated (r = 0.47), but more importantly, both components were highly correlated with item difficulty (r = -0.66 and r = 0.86). In other words, the classical factorial approach on dichotomous data could not distinguish substantive effects from methodological artifacts, a long-recognised issue even when using corrective correlation matrices (e.g., Bock et al., 1988; Kubinger, 2003).

In contrast to the above, three of the four analyses typically used in Rasch modelling suggested good data fit with the unidimensionality assumption: the Martin-Löf test (LR = 388.1, df = 574, p = 1.00), the non-parametric Ponocny *T1m* test (p = 1.00), and the proportion of response variance attributable to the Rasch model (33.2%, which is above the recommended lower practical threshold of 20.0%, according to Linacre, 2006). The fourth analysis, a principal component analysis on the residuals remaining after extracting the first component, yielded a result of 2.14, which, since it exceeded the 2.0 threshold, was considered grounds for additional item grouping review and the search for potential patterns (see Boone & Staver, 2020). Further scrutiny of individual items was also prompted by the results of other analyses characteristic of Rasch modelling: the local independence analysis, corrected item-total correlations, and the fit of items to the Rasch model using infit MSQ and standardised t-scores.

A content analysis following these psychometric procedures suggested that three items were indeed problematic. Item #22 was identified as redundant since it shared an identical solution rule and similar graphical shapes with item #27, as was indicated by the highest residual correlation among 1128 item pairs (Q_3 = .37). Items #33 and #42 showed poor model fit (#33, t = 7.24; #42, t = 8.51), accompanied by item-total correlations below .20. These were also the two items with loadings below .32 on the first unrotated factor. A content analysis revealed that these items had two plausible solutions. Thus, we excluded mentioned items from further analysis, with subsequent analyses conducted on the remaining 45 items (from now on PNT-45).

Improved psychometric indicators were obtained for all the previously mentioned criteria on the corrected scale (e.g., all items had loadings on the first unrotated factor greater than 0.32, the eigenvalue of the first component of the residuals was reduced to 1.98, and the proportion of variance explained by the Rasch model increased to 34.3%). Moreover, the internal consistency reliability of the summation score was high for both the classical and factor models (α = *0.89*, ω = *0.90*), as well as for the reliability within the Rasch model, R = *0.87* (see Anselmi, Colledani, & Robusto, 2019). However, one of the distinctive advantages of the Rasch model is its capacity to describe the informativeness of the test depending on the level of ability of the respondents. As seen in Figure 2, the distribution of correct answers is markedly asymmetric, and the test can be characterised as being easy. The results from our sample are almost identical to those obtained by Dautović and Borčić-Konjarek (1998) on a subsample of respondents aged 16 to 19. However, Figure 2 also shows that the test's informativeness is greatest for assessing low achievement, specifically within the range from -3 to 0 theta. From a generic perspective, which assumes a symmetric, ideally normal distribution of

Figure 2
Distribution of PNT-45 total scores (n = 721) and test information curve with standard errors (dashed line)



abilities, such a distribution might appear disadvantageous. However, in practice, this type of distribution is ideal for the purpose for which it is used by psychologists in evaluating required cognitive ability. Namely, the objective of such testing is to reliably detect drivers who lack the requisite ability, and thus the reliability must be highest around the cut-off values.

Encouraging results were also obtained through criterion validity analysis. To begin with, the results were in favour of the hypothesis of no gender differences (Males: M = 32.9, SD = 8.06, Females: M = 33.4, SD = 7.12; $t_w(665) = -0.93$, d = -0.07, BF_{01} = 7.69, p = .355). Additionally, no item showed significant indications of differential item functioning (DIF). Specifically, only four items had one significant indicator each, even though each was tested with four DIF tests (Mantel-Haenszel, standardised proportion test, logistic regression, and Raju's test). In contrast to gender, expected differences were found based on the type of school attended by the student. The type of school accounted for more than a quarter of the variance ($F(5, 715) = 50.69$, $\eta^2 = 0.26$, $BF_{10} > 1000$, $p < 0.001$), with three school clusters identified. The students from the gymnasium, music school, and electrotechnical school had comparable results (pooled M = 37.0, SD = 5.7), significantly higher than the students from technical schools (M = 31.6, SD = 6.2), as well as the students from polytechnic and agricultural schools (pooled M = 28.1, SD = 7.9).

Of particular interest is the finding that the students holding a driving license (n = 340, M = 34.9, SD = 6.6) achieved moderately higher scores than those without a license (n = 345, M = 31.4, SD = 8.2), a difference that was statistically significant ($t_w(656) = 6.14$, d = 0.47, $BF_{10} > 1000$, $p < .001$). On the other hand, a null correlation was found between PNT scores and attitudes towards risky driving behaviours (r = .00, BF_{01} = 21.24, p = .992), while a statistically inconclusive—but practically negligible—low positive correlation was observed with sensation seeking (r = .08, BF_{01} = 2.05, p = .031). These correlations are fully in line with meta-analytic findings regarding the relationships between impulsivity, sensation seeking, and related personality constructs with intelligence (Anglim et al., 2022; Stanek & Ones, 2023). Such finding corroborates the rationale behind regulations requiring that psychological evaluation of candidate drivers includes both cognitive ability and personality testing. This makes even more sense in light of recent empirical evidence suggesting that lower cognitive ability increases the likelihood of making erroneous decisions, but that it does not systematically increase risk-taking preferences (Mechera-Ostrovsky et al., 2022).

The final stages of the analysis were focused on norming. Thanks to Rasch modelling, it was possible to nonlinearly transform the raw scores into person parameter estimates along with their 95% confidence intervals. Due to the PNT's higher informativeness for the lower ability, the confidence intervals were narrower for below-average performance (e.g., for a score of 20, the confidence interval ranged from -2.56 to -1.15 θ) than for above-average results (e.g., for a score of 43, the confidence interval was from +0.63 to +3.34 θ). As a result, it was possible to create a table that also includes percentile ranks of raw scores and an indication of whether a given lower confidence interval falls into any of the areas of insufficient intelligence (defined as IQ thresholds of 70, 80, and 90 points, depending on the driving license category). Finally, using an automatic item selection algorithm (Becker et al., 2021) by maximising test information at the average ability level (θ = 0), it was possible to create a scale of only five medium-difficulty items, where a respondent needs to correctly answer three out of five items to reasonably eliminate the suspicion that they have below average performance.

Discussion and Conclusion

So, what is our study telling us about the state of the ancient PNT in the modern age? At first glance, the PNT appears to be holding up well. It is content-valid for assessing visually saturated information processing, and the corrected version shows good psychometric properties in terms of internal structure.

Reasonable criterion correlations were obtained, without indications of gender bias. In addition—and as we have already emphasised—scores were distributed in such a way that they were highly informative for detecting intellectual insufficiency. When we also consider the possibility of administering an optimised version that might save precious time to practitioners in approximately 50% of testing cases, can we then unreservedly recommend it for further use when evaluating drivers' cognitive ability?

Unfortunately, the answer is no. The first and most important reason for such an answer is that the PNT can be found on the internet—albeit through a thorough search. This automatically weakens the arguments for its legal use, even though the target groups are unlikely to prepare themselves for testing. The second reason is more of a paradigmatic dilemma about the comparative validity of using relatively general tests such as the PNT versus modern testing options, which include domain-specific testing in the simulators. Thirdly, it remains questionable, both from an ethical and functional perspective, whether norms derived from young adults should be applied to licence renewal candidates, particularly older experienced drivers. Finally, the negatively skewed distribution of the scores—while appropriate for screening out unsuitable driver candidates—limits the PNT's utility in other contexts such as recruitment or educational selection where the goal is typically to discriminate among candidates across the full range of ability. Despite these limitations, we believe that our study contributes to the discussion on the use of appropriate psychometric techniques for the validation and adaptation of tests such as the PNT.

For example, our review of regional literature has led us to the impression that Rasch modelling—only slightly younger than the PNT—is still a neglected approach compared to factor or classical modelling of ability tests, despite obvious advantages. Not only is Rasch modelling more consistently adapted to the statistical analysis of dichotomous data, but it also directly enables advanced options such as optimisation of administration through direct test shortening or computerised adaptive testing and norming that is particularly focused on specific ability levels. Rasch modelling—like other techniques belonging to the IRT family—places greater focus on item content and more thorough item analysis. Finally, if in the past the obstacle to using Rasch analysis was that it could only be done in commercial and rather hermetic software solutions, today almost all analyses can be done in open-source code using R packages, as was the case in our study.

In the end, we should answer the question posed in the title: whether using ancient tools provides benefits in the modern era. We are convinced that it does. For example, technological advancements and the current AI revolution allow for a more efficient generation of tests that would have items analogous to the PNT (e.g., Choi & Zhang, 2019; Gierl et al., 2015; Sayin et al., 2023). Considering all the psychometric advantages we have demonstrated in this paper, we believe it would be worthwhile to embark on creating a new version that would be administered in a computerised adaptive manner. That would also be an opportunity to upgrade the PNT by increasing its informativeness for higher visual-perceptual cognitive abilities, which might be beneficial for other selective purposes. All in all, old is not dead, but a thorough modernisation would be needed for the PNT to maintain its fitness and relevance in contemporary assessment. Before embarking on such a process, it is essential to clearly argue for the usefulness of developing such general-type tests and to ensure their ethical safeguarding in an age when it is hard to hide anything, so that we do not end up with yet another zombie wandering the Internet.

References

Anglim, J., Dunlop, P. D., Wee, S., Horwood, S., Wood, J. K., & Marty, A. (2022). Personality and intelligence: A meta-analysis. *Psychological Bulletin, 148*(5-6), 301–336. https://doi.org/10.1037/bul0000373

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology, 10*, 2714. https://doi.org/10.3389/fpsyg.2019.02714

Arthur, W., Barret, G. V., & Alexander, R. A. (1991). Prediction of vehicular accident involvement: A meta-analysis. *Human Performance, 4*(2), 89-105. https://doi.org/10.1207/s15327043hup0402_1

Becker, B., Debeer, D., Sachse, K. A., & Weirich, S. (2021). Automated test assembly in R: The eatATA package. *Psych, 3*(2), 96-112. https://doi.org/10.3390/psych3020010

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*(3), 387-409. https://doi.org/10.1348/096317905X26994

Bowen, L., Budden, S. L., & Smith, A. P. (2020). Factors underpinning unsafe driving: A systematic literature review of car drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 72*, 184-210. https://doi.org/10.1016/j.trf.2020.04.008

BIHAMK. (2018). Informacija o ukupnom broju aktivnih i proizvedenih/izrađenih vozačkih dozvola u BiH u periodu januar – decembar 2017 [Information on the total number of active and produced/issued driving licenses in BiH for the period January - December 2017]. https://bihamk.ba/assets/upload/informacija_o_vozackim_dozvolama%20-%20Copy%201.pdf

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280. https://doi.org/10.1177/014662168801200305

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods, 22*(3), 450-466. https://doi.org/10.1037/met0000074

Casutt, G., Martin, M., Keller, M., & Jäncke, L. (2014). The relation between performance in on-road driving, cognitive screening and driving simulator in older healthy drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 22*, 232-244. https://doi.org/10.1016/j.trf.2013.12.007

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Choi, J., & Zhang, X. (2019). Computerized item modeling practices using computer adaptive formative assessment automatic item generation system: A tutorial. *The Quantitative Methods for Psychology, 15*(3), 214-225. https://doi.org/10.20982/tqmp.15.3.p214

Collins, M. D. (2024). Train driver selection: The impact of cognitive ability on train driving performance. *International Journal of Selection and Assessment, 32*(2), 249-260. https://doi.org/10.1111/ijsa.12425

Costello, A. B., & Osborne, J. (2019). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), 7. https://doi.org/10.7275/jyj1-4868

Cyders, M. A., Littlefield, A. K., Coffey, S., & Karyadi, K. A. (2014). Examination of a short English version of the UPPS-P Impulsive Behavior Scale. *Addictive Behaviors, 39*(9), 1372-1376. https://doi.org/10.1016/j.addbeh.2014.02.013

Dautović, M., & Borcić-Konjarek, L. (1998). Use of the Purdue Non-verbal Test in a group of candidates for a driving license. *Arhiv za Higijenu Rada i Toksikologiju, 49*(2), 179-187.

Dugre, J. R., Giguere, C-É., Percie du Sert, O., Potvin, S., Dumais, A., & Consortium Signature (2019). The Psychometric Properties of a Short UPPS-P Impulsive Behavior Scale Among Psychiatric Patients Evaluated in an Emergency Setting. *Frontiers in Psychiatry, 10*, 1-9. https://doi.org/10.3389/fpsyt.2019.00139

Gierl, M. J., Ball, M. M., Vele, V., & Lai, H. (2015). A method for generating nonverbal reasoning items using n-layer modeling. In *Computer Assisted Assessment. Research into E-Assessment: 18th International Conference, CAA 2015* (pp. 12-21). https://doi.org/10.1007/978-3-319-27704-2_2

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

Kaber, D., Jin, S., Zahabi, M., & Pankok Jr, C. (2016). The effect of driver cognitive abilities and distractions on situation awareness and performance under hazard conditions. *Transportation Research Part F: Traffic Psychology and Behaviour, 42*, 177-194. https://doi.org/10.1016/j.trf.2016.07.014

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science, 45*(1), 106-110. https://doi.org/10.1002/0471264385.wei0205

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions, 20*(1), 1045. https://www.rasch.org/rmt/rmt201a.htm

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340-364. https://doi.org/10.1080/00273171.2011.564527

Lynam, D. R., Whiteside, S. P., Smith, G. T., & Cyders, M. A. (2006). *The UPPS-P: Assessing five personality pathways to impulsive behavior*. West Lafayette, IN: Purdue University. [Unpublished report]

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862. https://doi.org/10.3758/BRM.42.3.847

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20. https://doi.org/10.18637/jss.v020.i09

Mathias, J. L., & Lucas, L. K. (2009). Cognitive predictors of unsafe driving in older drivers: a meta-analysis. *International Psychogeriatrics, 21*(4), 637-653. https://doi.org/10.1017/S1041610209009119

McKnight, A. J., & McKnight, A. S. (1999). Multivariate analysis of age-related driver ability and performance deficits. *Accident Analysis & Prevention, 31*(5), 445-454. https://doi.org/10.1016/S0001-4575(98)00070-7

Mechera-Ostrovsky, T., Heinke, S., Andraszewicz, S., & Rieskamp, J. (2022). Cognitive abilities affect decision errors but not risk preferences: A meta-analysis. *Psychonomic Bulletin & Review, 29*(5), 1719-1750. https://doi.org/10.3758/s13423-022-02097-x

Peters, G. (2021). ufs: A collection of utilities for univariate frequency statistics and psychometrics [R package version 0.5.0]. CRAN. https://CRAN.R-project.org/package=ufs

Pravilnik o zdravstvenim uslovima koje mora ispunjavati vozač motornog vozila [Regulation on health requirements for motor vehicle drivers]. (2007). *Službeni glasnik Bosne i Hercegovine [Official Gazette of Bosnia and Herzegovina]*, 13/2007 & 89/2011.

Reger, M. A., Welsh, R. K., Watson, G., Cholerton, B., Baker, L. D., & Craft, S. (2004). The relationship between neuropsychological functioning and driving ability in dementia: a meta-analysis. *Neuropsychology, 18*(1), 85-93. https://doi.org/10.1037/0894-4105.18.1.85

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Robitzsch, A. (2020). sirt: Supplementary item response theory models [R package version 3.9-4]. CRAN. https://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis modules [R package version 3.5-19]. CRAN. https://CRAN.R-project.org/package=TAM

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of*

Applied Psychology, 88(6), 1068-1081. https://doi.org/10.1037/0021-9010.88.6.1068

Sayın, A., Bozdağ, S., & Gıerl, M. J. (2023). Automatic item generation for non-verbal reasoning items. *International Journal of Assessment Tools in Education, 10*(Special Issue), 132-148. https://doi.org/10.21449/ijate.1279120

Schultheis, M. T., Weisser, V., Ang, J., Elovic, E., Nead, R., Sestito, N., … & Millis, S. R. (2010). Examining the relationship between cognition and driving performance in multiple sclerosis. *Archives of Physical Medicine and Rehabilitation, 91*(3), 465-473. https://doi.org/10.1016/j.apmr.2009.11.026

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(1), 25-40. https://doi.org/10.1080/10705519609540027

Tendeiro, J. N., & Meijer, R. R. (2014). perfit: An R package for person-fit analysis in IRT. *Journal of Statistical Software, 74*(5), 1–27. https://doi.org/10.18637/jss.v074.i05

Torres Irribarra, D., & Freund, R. (2014). Wright Map: IRT item-person map with ConQuest integration. CRAN. https://CRAN.R-project.org/package=WrightMap

Triffin, J., Grubner, A., & Inaba, K. (1958). *Purdue non-language test: Preliminary Manual*. West Lafayette: Occupational Research Center, Purdue University.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327. https://doi.org/10.1007/BF02293557