A gamified Faux Pas Test

Comparing psychometric properties to the pen-and-paper version in a Dutch population

Poos, Jackie M.; Zinzen, Indy; Kalisvaart, Max; Assendelft, Linde; Marticic Giljevic, Karla; Ruitenberg, Marit; Bidarra, Rafael; van den Berg, Esther

DOI
10.1111/jnp.70015

Publication date
2025

Document Version
Final published version

Published in
Journal of Neuropsychology

Received: 30 June 2025 | Accepted: 28 September 2025

DOI: 10.1111/jnp.70015

the british psychological society
promoting excellence in psychology

BRIEF COMMUNICATION

A gamified Faux Pas Test: Comparing psychometric properties to the pen-and-paper version in a Dutch population

Jackie M. Poos[1] | **Indy Zinzen**[1,2] | **Max Kalisvaart**[1,2] | **Linde Assendelft**[1] | **Karla Marticic Giljevic**[1] | **Marit Ruitenberg**[2,3] | **Rafael Bidarra**[4] | **Esther van den Berg**[1]

[1]Department of Neurology and Alzheimer Center Erasmus MC, Erasmus MC University Medical Center, Rotterdam, The Netherlands

[2]Department of Health, Medical and Neuropsychology, Leiden University, The Netherlands

[3]Leiden Institute for Brain and Cognition, Leiden, The Netherlands

[4]Department of Computer Science, Delft University of Technology, Delft, The Netherlands

Correspondence
Jackie M. Poos, Dr. Molewaterplein 40, 3015GD Rotterdam, The Netherlands.
Email: j.m.poos@erasmusmc.nl

Funding information
Alzheimer Nederland; Association for Frontotemporal Degeneration

Abstract

Psychometric properties of Tommy's Quest (TQ), a novel serious game to evaluate Theory of Mind (ToM) and the pen-and-paper Faux Pas Test (FPT) were assessed. Results from 67 cognitively unimpaired individuals indicated that TQ had adequate construct validity, internal consistency and test–retest reliability. Participants performed worse on TQ compared to the FPT, suggesting greater sensitivity to subtle deficits. These findings support serious games like TQ as a promising tool for ToM assessment, highlighting the need for clinical validation.

KEYWORDS

dementia, diagnosis, neuropsychological assessment, serious games, social cognition, theory of mind

INTRODUCTION

Theory of Mind (ToM), the ability to attribute beliefs, desires and intentions to others, is a core component of social cognition that is frequently affected in neurocognitive disorders (Dodich et al., 2025). Despite its recognized importance, the clinical assessment of ToM remains fragmented and inconsistent. Neuropsychologists have low confidence in the validity of available measures, as existing measures are often limited by inadequate psychometric properties (Quesque et al., 2024; Quesque & Rossetti, 2020).

Recent work has emphasized conceptual and methodological challenges in ToM assessment, particularly the lack of specificity of many classical tasks (Quesque & Rossetti, 2020). A major concern is that available measures often fail to meet core criteria for ToM: their demands can be explained by lower-level processes such as attention, visual discrimination or emotion recognition (the *mentalizing*

criterion), and/or they do not require a clear distinction between one's own and another's mental state (the *non-merging criterion*). Quesque and Rossetti (2020) argue that addressing these limitations requires the development of measures with stronger ecological validity. Most current ToM tasks, however, rely on static photographs, written stories or dynamic video clips in which participants remain passive observers (Oker, 2022; Quesque & Rossetti, 2020). Research has shown that such paradigms are not consistent with what happens in everyday life (Quesque & Rossetti, 2020) and has highlighted the importance of interactive and immersive environments for investigating social cognition (Oker, 2022). Serious games represent a promising avenue to address this gap. They can simulate real-world conditions, incorporate distractors absent in standard clinical settings and provide interactive environments that elicit more naturalistic emotional and behavioural responses (Martínez-Pernía et al., 2025).

To address this gap, we developed Tommy's Quest (TQ), a serious game designed to assess ToM using faux pas scenarios embedded within a gamified narrative (Bekooy et al., 2023). The primary aim of the present study was to evaluate the psychometric properties of TQ in cognitively unimpaired adults and to compare these properties with those of a classical pen-and-paper measure of ToM based on faux pas scenarios, the Faux Pas Test (FPT) (Baron-Cohen et al., 1999).

METHODS

Participants

Cognitively unimpaired participants were recruited via community homes and friends and families of the researchers. All assessments were conducted at their homes. Inclusion criteria were: (i) above 18 years old; (ii) Dutch native speaker. Exclusion criteria were: (i) a medical condition known to affect cognition. All participants gave written informed consent. Ethical approval was obtained from Leiden University.

Measures

Tommy's quest

TQ (see Figure 1) (Bekooy et al., 2023) is a serious game assessing a person's ability to understand if something awkward or insulting has been said unintentionally during a two-person social interaction. In contrast to the original FPT, participants are not merely reading or hearing about a social interaction, but are immersed in the storyline in which their own game character has social interactions with other characters in the game. In TQ, participants follow Tommy, a boy who has lost his cat. To find his cat, Tommy goes on an adventure around the neighbourhood, interacting with other characters. The game consists of several episodes: a supermarket, a playground, a school and a public street. Each episode consists of several dialogues (i.e. test-items). After each dialogue, participants are asked to judge whether someone said something inappropriate, and they can select 'yes' or 'no' from multiple-choice options, followed by a multiple-choice question offering three options that explain what was unintentionally said. Scores are derived based on correct answers, with a maximum of 18 points: 6 for recognizing non-faux pas scenarios, 6 for faux pas scenarios and 6 for correctly explaining why a faux pas occurred. Administration time is 15 minutes. A detailed description of TQ's development can be found in the Supporting Information and Bekooy et al. (2023).

Other neuropsychological tests

A neuropsychological test battery, including TQ and the FPT, was administered to evaluate language, mental processing speed, executive function, memory and social cognition (see Table 1). We also used

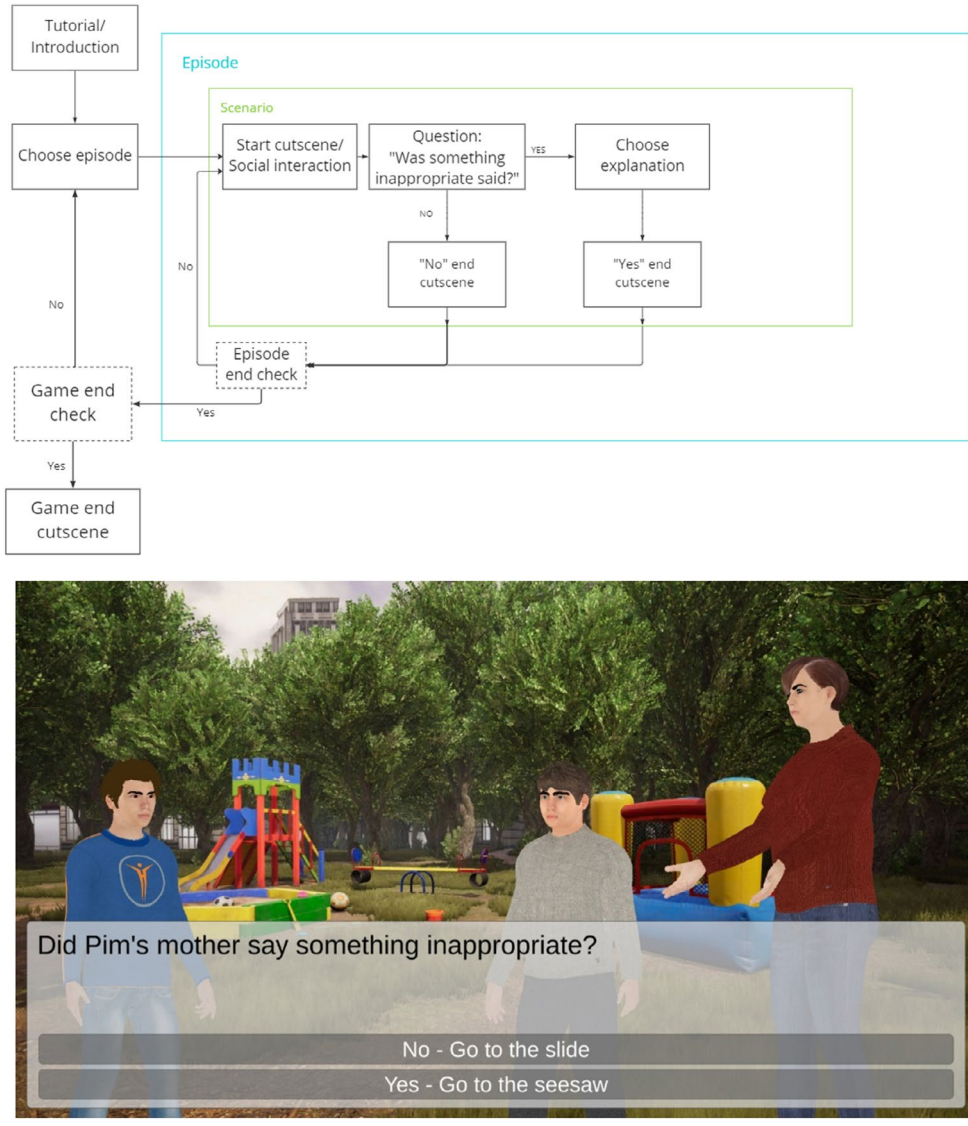

FIGURE 1 The game loop and a screenshot of Tommy's Quest (Bekooy et al., 2023).

the User Experience Questionnaire (UEQ), consisting of 26 opposing game characteristics that were rated on a 7-point Likert scale. Total administration time was 120 minutes.

Statistical analysis

Participants with outliers ≥3 tests were removed. Missing data were imputed by replacing them with the mean of the available data. Spearman correlational analyses investigated the relationship between TQ, other neuropsychological tests, age, education level and test–retest reliability. Sex differences were assessed with independent-samples Mann–Whitney U tests. Internal consistency was determined by calculating Cronbach's alpha. Total scores on TQ were transformed to align with the scale of the FPT

as follows: (total score/18)*20. Performance on TQ was compared to the FPT with a Wilcoxon signed-rank test. All analyses were conducted in IBM SPSS Statistics 28.01.0.

RESULTS

Demographics

One outlier was excluded, resulting in a sample of sixty-seven participants with 58.2% being female (for descriptives, see Table 1). TQ scores did not differ between sexes ($U(66) = .5$, $p = .6$) and were not related to age ($\rho = -.2$, $p = .2$), but showed a significant correlation with education level ($\rho = .3$, $p < .01$). The pen-and-paper FPT did not differ between sexes ($U(66) = -.6$, $p = .6$) but showed significant correlations with age ($\rho = -.4$, $p < .01$) and education level ($\rho = .5$, $p < .01$). Table S1 demonstrates all correlation coefficients.

TABLE 1 Demographics and neuropsychological test scores.

	Total	Male	Female
n	67	28	39
Age	51.3 (17.7)	51.8 (17.8)	51.0 (17.9)
Education level[a,b]	5.8 (1.0)	6.2 (.9)	5.5 (1.1)
MoCa (max = 30)	26.3 (2.1)	26.3 (2.0)	26.4 (2.2)
Tommy's Quest (max = 18)	14.2 (2.6)	14.0 (2.8)	14.4 (2.4)
Faux pas items (max = 12)	8.8 (2.4)	8.5 (2.7)	8.8 (2.1)
Non-Faux pas items (max = 6)	5.5 (1.1)	5.5 (.5)	5.5 (.8)
Faux Pas Test (max = 20)	18.0 (2.1)	18.1 (2.3)	18.0 (1.9)
Animal fluency	25.6 (5.5)	25.9 (5.9)	25.3 (5.2)
TMT a, seconds (max = 300)	35.5 (16.2)	35.9 (18.9)	35.3 (14.2)
TMT b, seconds (max = 300)	76.8 (39.9)	80.0 (41.0)	74.6 (39.6)
RAVLT (max = 48)	47.6 (9.7)	45.5 (11.0)	49.1 (8.5)
FEEST (max = 60)[b]	47.0 (5.3)	45.3 (5.7)	48.2 (4.7)
ERT (max = 96)	56.4 (9.9)	54.1 (10.6)	58.0 (9.9)
Hinting task (max = 12)	11.3 (1.0)	11.2 (1.2)	11.4 (.7)
SET (max = 18)	16.8 (1.3)	16.8 (1.0)	16.8 (1.5)
Cartoons task (max = 12)[b]	7.8 (2.3)	8.8 (2.3)	7.1 (2.2)
SNQ (max = 22)	19.2 (1.3)	19.3 (1.4)	19.1 (1.4)
RSMS (max = 78)	43.3 (7.1)	44.6 (5.9)	42.3 (7.8)
MBI (max = 84)	54.8 (6.9)	54.0 (6.3)	55.3 (7.3)
TAS20 (max = 100)	46.5 (11.8)	47.0 (11.0)	46.1 (12.4)

Note: All data are presented as mean (SD) unless otherwise stated. Data was available for all participants except: FEEST (65/67), ERT (66/67), Hinting Task (66/67), SET (63/67), Cartoons Task (57/67).

Abbreviations: ERT, Emotion Recognition Test; FEEST, Facial Expressions of Emotions Test and Stimuli; MBI, Moral Behavioural Inventory; MoCa, Montreal Cognitive Assessment; RAVLT, Rey Auditory Verbal Learning Test; RSMS, Revised Self-Monitoring Scale; SET, Story-based Empathy Task; SNQ, Social Norms Questionnaire; TAS20, Toronto Alexithymia Scale 20 items; TMT, Trail Making Test.

[a]Education level according to the Dutch Verhage system.

[b]Significant difference between sexes at $p < .05$.

Convergent and divergent validity

Significant correlations were observed between TQ and the FPT ($\rho = .5$, $p < .01$), Hinting Task ($\rho = .4$, $p < .01$), SET ($\rho = .3$, $p = .02$), Cartoons task ($\rho = .3$, $p < .01$), ERT ($\rho = .3$, $p = .02$), TAS20 ($\rho = -.3$, $p = .04$), TMTb ($\rho = -.3$, $p = .01$) and animal fluency ($\rho = .3$, $p = .01$).

The pen-and-paper FPT significantly correlated with FEEST ($\varrho = .4$, $p < .01$), ERT ($\varrho = .4$, $p < .01$), Hinting task ($\varrho = .3$, $p = .03$), Cartoons task ($\varrho = .4$, $p < .01$), SNQ ($\varrho = .4$, $p < .01$), MoCa ($\varrho = .5$, $p < .01$), TAS20 ($\varrho = -.4$, $p < .01$), TMTa ($\varrho = -.4$, $p < .01$), TMTb ($\varrho = -.4$, $p < .01$) and RAVLT ($\varrho = .4$, $p < .01$).

Internal consistency

For TQ, Cronbach's alpha was .7 for the FP items. The corrected item-total correlation ranged from $\alpha = .2$ to $\alpha = .4$, with nine out of twelve items having $\alpha > .3$. Item-total statistics indicated that the removal of any item would result in equal or lower internal consistency. Regarding the non-FP items, Cronbach's alpha was .2. The corrected item-total correlation ranged from 0 to .2.

For the FPT, Cronbach's alpha was .6 for the FP items. The corrected item-total correlation ranged from $\alpha = .1$ to $\alpha = .4$, with five out of ten items having $\alpha > .3$. Item-total statistics indicated that the removal of any item would result in equal or lower internal consistency. Regarding the non-FP items, Cronbach's alpha was .3.

Test–retest reliability

The correlation between the two time points ($n = 10$) was at trend level for TQ ($\rho = .6$, $p = .08$), but low for the FPT ($\rho = .2$, $p = .5$).

Comparison to the Faux Pas Test

Performance on TQ (Med = 14.0, SE = .3) was significantly poorer than on the pen-and-paper FPT (Med = 18.0, SE = .3) ($W(66) = -5.4$, $p < .01$). The variance in test scores is illustrated in Figure S1.

User experience

Overall, participants rated the game favourably, particularly regarding comprehensibility and ease of use, while ratings for creativity, excitement and motivation could be improved (see Table 2).

DISCUSSION

This study demonstrated adequate psychometric properties of TQ, a digital neuropsychological game, suggesting that it is a valid measure of ToM. TQ was rated positively by users, showed comparable internal consistency to the FPT and outperformed the FPT in terms of test–retest reliability. Performance on TQ was correlated with all other tests measuring ToM, as well as all tests measuring perception and attribution, suggesting high convergent validity. The FPT correlated with the Cartoons and Hinting tasks, all tests measuring perception and attribution, and the SNQ, but not with the SET.

Considering divergent validity, TQ was associated with two tests measuring executive function and language abilities. This might be explained by the fact that TQ relies on some degree of language

TABLE 2 Mean and standard deviations for the User Experience Questionnaire on Tommy's Quest.

Statement	Mean (SD)
Unpleasant – Pleasant	5.3 (1.4)
Incomprehensible – Comprehensible	5.8 (1.4)
Creative – Boring	4.0 (1.9)
Easy to learn – Difficult to learn	2.5 (1.9)
Valuable – Inferior	3.8 (1.3)
Boring – Exciting	4.4 (1.2)
Uninteresting – Interesting	4.6 (1.4)
Unpredictable – Predictable	4.9 (1.6)
Fast – Slow	4.0 (1.8)
Original – Conventional	3.5 (1.9)
Hindering – Supportive	4.9 (1.2)
Good – Bad	3.2 (1.5)
Complex – Simple	5.7 (1.4)
Repulsive – Attractive	4.7 (1.3)
Common – New	4.8 (1.5)
Unpleasant – Pleasant	5.3 (1.3)
Familiar – Unfamiliar	3.1 (1.5)
Motivating – Demotivating	3.4 (1.7)
As expected – Not as expected	3.4 (1.6)
Inefficient – Efficient	5.0 (1.3)
Clear – Confusing	2.7 (1.6)
Impractical – Practical	5.6 (1.0)
Orderly – Messy	2.5 (1.4)
Attractive – Unattractive	3.5 (1.6)
Nice – Unkind	2.5 (1.3)
Conservative – Innovative	4.4 (1.6)

Note: Lower scores reflect a tendency towards the left on the continuum, and higher scores a tendency towards the right on the continuum.

processing and executive functioning to navigate the game environment and understand the scenarios presented. Indeed, prior research demonstrated the involvement of executive functions in playing video games (Chen & Hsieh, 2018). In contrast, the FPT exhibited broader correlations spanning global cognitive functioning, attention, processing speed, executive function, language and episodic memory. This suggests the FPT may tap into a wider range of cognitive domains, potentially impacting its specificity as a ToM measure. One reason for this could be that the FPT requires reading and/or listening to longer narratives without visual/audio stimuli, thereby placing a greater demand on memory and attention processes.

This also raises questions about the content validity of TQ and the FPT. Our findings showed broad correlations between the FPT and multiple cognitive domains, including global cognitive status, suggesting that it does not consistently fulfill *the mentalizing criterion* (Quesque & Rossetti, 2020). TQ also relied to some extent on executive and language abilities and therefore cannot be considered 'process pure'. However, by embedding faux pas scenarios in an interactive and ecologically valid game context, TQ may more directly elicit mental state attribution. In this sense, it represents a step towards tasks that align more closely with the theoretical requirements of ToM assessment.

TQ and the FPT were associated with education level, which is consistent with previous research showing that higher educational attainment mitigates age-related decline in ToM tasks (Li et al., 2013).

The FPT was also associated with age, likely reflecting the test's reliance on other cognitive processes that are known to decline with aging. Notably, sex did not affect performance on either TQ or the FPT, contrary to literature (Greenberg et al., 2023). This may partly stem from the fact that women in the study had lower educational levels compared to their male counterparts, which influenced performance on ToM.

Lastly, compared to the FPT, performance on TQ was significantly poorer, making it plausible that TQ is more challenging and potentially more sensitive to subtle social cognitive deficits. The next critical step is to validate TQ in clinical populations.

CONCLUSION

This study demonstrated that a serious game adaptation of a widely used ToM test is a valid and reliable measure of social cognition, with evidence that it may be more sensitive than the classical pen-and-paper FPT. These findings highlight the promise of serious games as tools for advancing the assessment of social cognition, though further validation in clinical populations remains essential.

AUTHOR CONTRIBUTIONS

Jackie M. Poos: Conceptualization; investigation; funding acquisition; writing – original draft; methodology; validation; visualization; writing – review and editing; formal analysis; project administration; data curation; supervision; resources. **I. Zinzen:** Investigation; writing – review and editing; methodology; visualization; formal analysis. **M. Kalisvaart:** Investigation; methodology; visualization; writing – review and editing; formal analysis; project administration. **Linde Assendelft:** Conceptualization; writing – review and editing; methodology; supervision. **K. Marticic Giljevic:** Conceptualization; writing – review and editing. **Marit Ruitenberg:** Conceptualization; methodology; writing – review and editing; supervision. **R. Bidarra:** Software; supervision; resources; writing – review and editing. **E. van den Berg:** Conceptualization; writing – review and editing; methodology; investigation; resources.

ACKNOWLEDGEMENTS

The authors acknowledge the assistance of an AI language model in the refinement of this manuscript. We thank Mark Bekooy, Dan Berendsen, Martin Dierikx, Rolf Piepenbrink and Jan-Willem van Rhenen from Delft University of Technology for the development of the TQ software. This work was funded by Alzheimer Nederland and the Association for Frontotemporal Dementias.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data are available upon request with the corresponding author.

ORCID

Jackie M. Poos https://orcid.org/0000-0001-8843-7247
Linde Assendelft https://orcid.org/0009-0001-4870-5371
Marit Ruitenberg https://orcid.org/0000-0001-7435-3229
Rafael Bidarra https://orcid.org/0000-0003-4281-6019

REFERENCES

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syn drome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*, 407–418.

Bekooy, M., Berendsen, D. D., Dierikx, M., Piepenbrink, R., van Rhenen, J. W. & Bidarra, R. (2023). A theory-of-mind game for the early detection of frontotemporal dementia. In *International Conference on Interactive Digital Storytelling*, LNCS 14384, pp. 137–145. Springer.

Chen, Y. Q., & Hsieh, S. (2018). The relationship between internet-gaming experience and executive functions measured by virtual environment compared with conventional laboratory multitasks. *PLoS One*, *13*(6), e0198339.

Dodich, A., Panzavolta, A., Funghi, G., Meli, C., Festari, C., Chatzikostopoulos, T., Chicherio, C., Clarens, F., de Oliveira, F. F., Filardi, M., Ibanez, A., Invernizzi, L., Lebouvier, T., Logroscino, G., MacPherson, S. E., Manca, R., Marra, C., Matias-Guiu, J. A., Montembeault, M., … SIGNATURE Consortium. (2025). International Consensus for the Assessment of Social Cognition in Neurocognitive Disorders: Framework Definition and Clinical Recommendations of the SIGNATURE Initiative.

Greenberg, D. M., Warrier, V., Abu-Akel, A., Allison, C., Gajos, K. Z., Reinecke, K., & Baron-Cohen, S. (2023). Sex and age differences in "theory of mind" across 57 countries using the English version of the "Reading the mind in the eyes" test. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(1), e2022385119.

Li, X., Wang, K., Wang, F., Tao, Q., Xie, Y., & Cheng, Q. (2013). Aging of theory of mind: The influence of educational level and cognitive processing. *International Journal of Psychology*, *48*(4), 715–727.

Martínez-Pernía, D., Olavarría, L., Fernández-Manjón, B., Cabello, V., Henríquez, F., Robert, P., & Slachevsky, A. (2025). The limitations and challenges in the assessment of executive dysfunction associated with real-world functioning: The opportunity of serious games. *Applied Neuropsychology. Adult*, *32*(2), 557–573.

Oker, A. (2022). Embodied social cognition investigated with virtual agents: The infinite loop between social brain and virtual reality. *Frontiers in Virtual Reality*, *3*, 962129.

Quesque, F., Nivet, M., Etchepare, A., Wauquiez, G., Prouteau, A., Desgranges, B., & Bertoux, M. (2024). Social cognition in neuropsychology: A nationwide survey revealing current representations and practices. *Applied Neuropsychology. Adult*, *31*(4), 689–702.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384–396.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1.

How to cite this article: Poos, J. M., Zinzen, I., Kalisvaart, M., Assendelft, L., Marticic Giljevic, K., Ruitenberg, M., Bidarra, R., & van den Berg, E. (2025). A gamified Faux Pas Test: Comparing psychometric properties to the pen-and-paper version in a Dutch population. *Journal of Neuropsychology*, *00*, 1–8. https://doi.org/10.1111/jnp.70015