2023
Volume 2
Article 13
https://doi.org/10.59453/ll.v2.13

Is artificial intelligence more creative than humans? ChatGPT and the Divergent Association Task

David Cropley

Centre for Change and Complexity in Learning, University of South Australia

A fundamental premise of the future of work is that AI will replace people in many cognitive and physical tasks, leaving creativity as a core, human 21st century skill. However, the recent launch of generative AI (especially ChatGPT) has seen many claims that AI is creative. If true, then the foundation of future human work, and education, is under threat. To examine claims of AI creativity, this research applied a test of verbal divergent thinking – the Divergent Association Task – to two versions of ChatGPT (GPT3.5 and GPT4). The results are reported and compared to a large human baseline. While both forms of ChatGPT show a capacity for verbal divergent production that exceeds human means, a range of factors call into question the "creativity" of generative AI.

Keywords: artificial intelligence, ChatGPT, creativity, divergent thinking

Corresponding author: David Cropley, david.cropley@unisa.edu.au

Recommended citation: Cropley, D. (2023). Is artificial intelligence more creative than humans? ChatGPT and the divergent association task. *Learning Letters*, *2*, Article 13. https://doi.org/10.59453/ll.v2.13

Introduction

The rise of Artificial Intelligence (AI), especially the recent emergence of OpenAI's large language model ChatGPT, has reinvigorated the 'future of work' debate. Even only recently (e.g., Cropley et al., 2022; Frey & Osborne, 2017) we were reassured that AI would replace humans only in jobs that were characterised by *predictable*, algorithmic cognitive and physical labour (e.g., bookkeeping, editing) leaving humans free to focus on unpredictable, non-algorithmic cognitive and physical work reliant on soft skills such as emotional intelligence, complex problem solving, and *creativity*. Thus, AI held no fears for us, provided we prepared ourselves appropriately. This has been the catalyst for a renewed focus on soft skills in education systems around the world (e.g., Succi & Canovi, 2020).

However, the emergence of ChatGPT, among a suite of large language models, including "Bard" and "Claude", has ignited a new debate, threatening to disrupt the balance between humans and AI central to the notion of the future of work.[1] Many claims have emerged attributing considerable creative ability to AI (Du Sautoy, 2020). This includes not just the verbal abilities of ChatGPT (Henriksen et al., 2023), but also the image-generation abilities of DALL-E 2 (Kirkpatrick, 2023), and even artificial musical creativity (Gioti, 2021). Notwithstanding the considerable misinterpretation, and even misuse, of the term "creativity" that confounds this debate (Cropley et al., 2019), a simple way to explore this hypothesis – AI is more creative than humans – is to administer a robust, validated and normed creativity test to the AI.

[1] "Bard" was released in March 2023 by Google, while "Claude" was released at a similar time by Anthropic.

There is a wide variety of tests that examine different facets of the construct “creativity”, from the broad and holistic, to the narrow and focused. The Test of Creative Thinking: Drawing Production (TCT-DP: Urban & Jellen, 1996), for example, examines elements of personality, cognition and creative output. By comparison, the Alternate Uses Test (AUT: Torrance, 1999), addresses specific elements of creative cognition (i.e., divergent thinking). These tests also use a variety of media, from the figural (the TCT-DP) to the verbal (the AUT). The fact that ChatGPT is a large *language* model AI means that a verbal test is the obvious and salient choice.

However, a limitation of many creativity tests is scoring, with the majority relying on human raters. Although many tests, such as the TCT-DP, have detailed scoring criteria for use by trained raters, subjectivity cannot be ruled out. This is reflected in the fact that interrater agreement on the TCT-DP is typically around .90. Divergent thinking tests such as the AUT rely on apparently more objective criteria (e.g., fluency, flexibility, originality), however, these tests are typically slow, and remain vulnerable to subjectivity.

Fortunately, recent advances in AI have addressed these scoring limitations, with machine learning approaches demonstrating a strong capacity to score figural (e.g., Cropley & Marrone, 2022) and written/verbal (Marrone et al., 2022) creativity tests. In particular, Olson et al. (2021) have developed the Divergent Association Task (DAT), using the concept of semantic distance (see Beaty & Johnson, 2021), to create a robust, automated test of verbal divergent thinking that is ideally suited to ChatGPT. Insofar as semantic distance is an acceptable proxy for divergent thinking (Olson et al., 2021 show strong correlations between the DAT and both flexibility and originality on the AUT) and noting that divergent thinking is an important, but not sole, indicator of creativity (e.g., Plucker, Makel & Qian, 2019; Runco & Acar, 2019) the DAT is used here as a valid proxy of creativity.

Therefore, to explore the hypothesis that ChatGPT is more creative than humans, two versions of the model (GPT3.5 and GPT4) were given the DAT with results compared to the human norms reported by Olson et al. (2021). The method used to conduct this study, the results obtained, and a discussion of these are set out in the following sections.

Method

The standard DAT instruction (Olson et al., 2021, p. 5) is as follows:

> *Please enter 10 words that are as different from each other as possible, in all meanings and uses of the words.*
>
> *Rules:*
>
> 1. *Only single words in English.*
> 2. *Only nouns (e.g., things, objects, concepts).*
> 3. *No proper nouns (e.g., no specific people or places).*
> 4. *No specialised vocabulary (e.g., no technical terms).*
> 5. *Think of the words on your own (e.g., do not just look at objects in your surroundings).*

These instructions were entered in the ChatGPT user interface with one modification: “please enter” was replaced with “give me”. Multiple sets of words were collected from ChatGPT across a number of different sessions.

Each set of ten words generated by ChatGPT (both GPT3.5 and GPT4) was copied and saved in a master spreadsheet (available on request). To score these responses, each set of words was entered into the DAT website (https://www.datcreativity.com/task) which returned two values: (a) the raw score (ranging from 0 to 200) and (b) the percentile rank of

that score, comparing the raw score to the large sample of individuals who have completed the DAT (see Olson et al., 2021).

The decision to collect approximately 100 responses (i.e., 100 sets of 10 words) each from GPT3.5 and GPT4 was determined by normal power analysis considerations (i.e., Cresswell, 2005, recommends statistical significance set at p = .05, power criterion set at .80 and effect size = .40, sample sizes of 100).

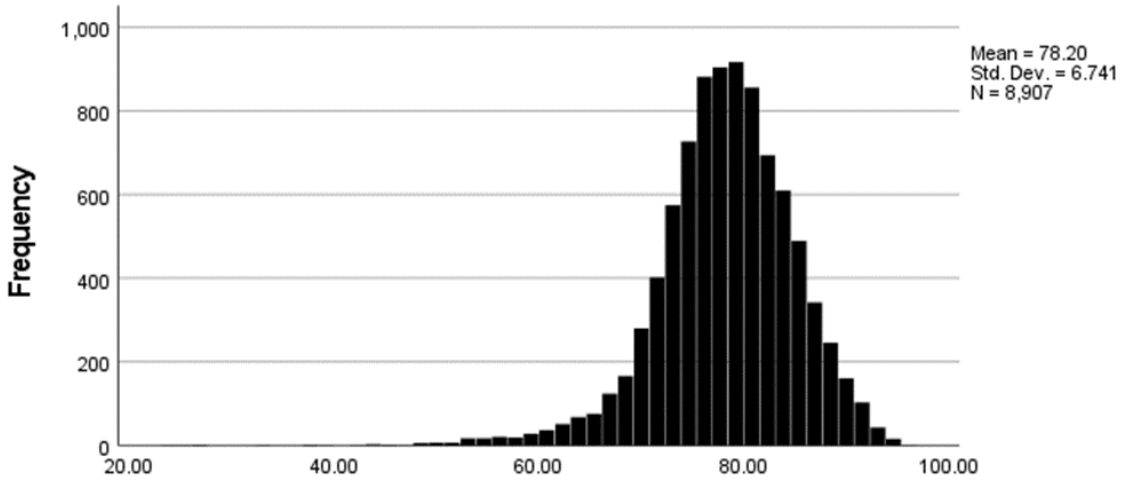
Results

The DAT raw score data reported by Olson et al. (2021) establishes a baseline for the comparison of ChatGPT (GPT3.5 and GPT4) data with a large human sample (i.e., a norm). The DAT norm data (Table 1), though negatively skewed (-.95) and leptokurtic (3.62), nevertheless can be regarded as normally distributed (Figure 1) and therefore suitable for further parametric analyses.

Table 1: Descriptive data, DAT scores (Norm, GPT3.5, GPT4)

	N	Mean	SD	Min	Max	Skewness	Kurtosis
DAT (Norm) Raw Score	8907	78.20	6.74	23.85	95.74	-.94	3.62
DAT (Norm) Percentile Rank	8907	50.00	28.87	.01	100.00	.00	-1.20
DAT (GPT3.5) Raw Score	102	80.89	4.36	69.28	89.78	-.75	.19
DAT (GPT3.5) Percentile Rank	102	64.93	21.75	9.58	96.11	-.91	.02
DAT (GPT4) Raw Score	102	85.28	4.01	74.69	95.69	.04	.93
DAT (GPT4) Percentile Rank	102	82.54	13.94	31.03	99.60	-1.32	1.87

Figure 1: DAT raw score baseline (norm) distribution



The data collected for ChatGPT responses on the DAT were also normally distributed

(Table 1)[2] and therefore suitable also for further parametric analyses. Percentile rank scores for the ChatGPT data (Table 1) were also calculated for the purposes of aiding in the comparison of ChatGPT and human responses on the DAT.

A one-way analysis of variance (ANOVA) was conducted to explore the creativity (divergent thinking) of ChatGPT relative to a human sample, as measured by the DAT. Three groups were utilised in this study (Group 1: human respondents; Group 2: ChatGPT (GPT3.5); Group 3: ChatGPT (GPT4)). There was a statistically significant difference at the level $p < .001$ in DAT scores for the three groups: F (2, 9108) = 63.99, $p < .001$. Despite reaching statistical significance, the actual difference in mean scores between the groups was small. The effect size, calculated using eta squared, was .01. Post-hoc comparisons using the Tukey HSD test indicated that the mean scores for Group 1 (M = 78.20, SD = 6.74), Group 2 (M = 80.89, SD = 4.36) and Group 3 (M = 85.28, SD = 4.01) were all significantly different from each other.

Discussion

At first glance, the performance of ChatGPT on the Divergent Association Task (DAT) seems impressive. The mean DAT response of GPT3.5, for example, is higher than 64.93% of human responses (that is, 0.4 of a standard deviation above the norm). More significantly, the mean DAT response of GPT4 is higher than 82.54% of human responses (1.04 standard deviations above the norm). The latter, certainly, suggests that GPT4, while not at the level of Bloom's (1984) two-sigma problem, is superior to most humans. However, before the hypothesis – AI is more creative than humans – can be supported, several important factors need further consideration.

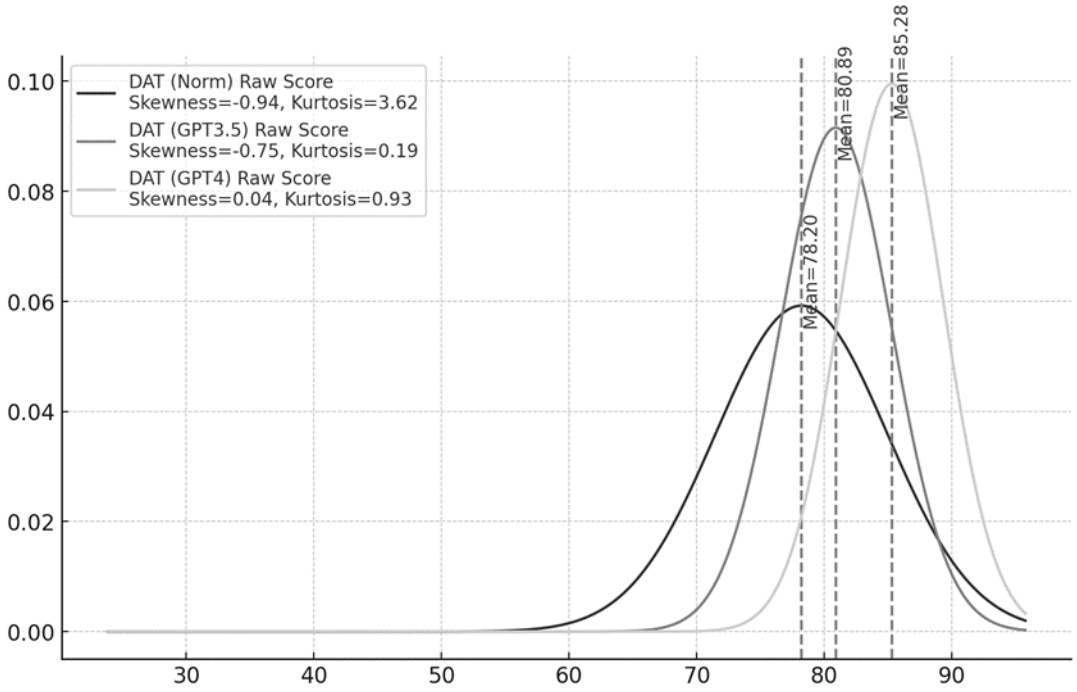
First, while the statistical test (ANOVA) indicates that both GPT3.5 and GPT4 had statistically significant, higher mean scores than the human norm sample, the critical consideration is the effect size (eta squared). This was small, indicating that, while a difference does exist, it is likely of little practical importance. On this basis, only qualified support can be given to the hypothesis.

The question of the *practical importance* of the differences in DAT scores, between ChatGPT and humans, can be understood better by superimposing the probability density functions of each dataset on each other (Figure 2).

What Figure 2 helps to illustrate is that, although the mean DAT score for GPT3.5 (80.89) sits approximately at the 65th percentile relative to the norm, there remains a significant likelihood (approximately 35%) that any given human response on the DAT will be higher than the GPT3.5 mean. Conversely, a significant proportion of GPT3.5 responses (approx. 25%) lie below the mean of the norm.

[2] Values of Skewness and Kurtosis (Table 1) for all variables fall within accepted limits for normality. The Kolmogorov-Smirnov test was significant (i.e., normal) for the GPT3.5 and GPT4 samples, while inspection of histograms (e.g., Figure 1), Normal Q-Q and Detrended Normal Q-Q plots support normality for all variables (Tabachnick & Fidell, 2013).

Figure 2: Comparing probability density functions of DAT raw scores (Norm, GPT3.5, GPT4)



GPT4, by comparison, performs somewhat better than GPT3.5 relative to the norm. The mean DAT score for GPT4 (85.28) sits approximately at the 83rd percentile relative to the norm distribution. Therefore, only approximately 17% of human responses will be higher than the GPT4 mean. While this is more impressive, there remain, nevertheless, approximately 5% of GPT4 responses that fall below the mean of the norm. It seems reasonable, therefore, to conclude that while GPT3.5 is, on average, somewhat higher on the DAT than a human respondent, this difference is not large enough or sufficiently frequent, to warrant replacing a human, unless there are other factors at play (such as the speed of the response). Conversely, it appears that the performance of GPT4 on the DAT, relative to humans, is sufficient to warrant replacing a human. However, this may also be tempered by other considerations. Broadly speaking, it appears that there is some justification for claiming that GPT4 outperforms humans on the DAT. However, returning to the limitation that divergent thinking is an important (but not the only) component of creativity, this result provides only qualified support for the claim that ChatGPT (in the form of GPT4) is more creative than humans.

These statistics also highlight some weaknesses in ChatGPT, relative to divergent thinking, that further assist in testing the hypothesis. Put simply, if ChatGPT is *creative* (insofar as it performs well in verbal divergent production), it is unreliably so. The substantial overlap in the distributions means that, while any given GPT4 response is likely to be higher than a human response, there is no guarantee that this will be the case. A high-performing and consistent human, therefore, is likely to be a far more dependable source of verbal divergent thinking than GPT4.

A second weakness of both GPT3.5 and GPT4 is also evident from the data. Across the 102 sets of 10 responses produced by both models, many words are repeated (Table 2). While the level of repetition improved in GPT4, compared to GPT3.5, it remains high enough to cause concerns.

Table 2: Repetition of words in the datasets

GPT3.5			GPT4		
Word	**Frequency**	**Percentage**	**Word**	**Frequency**	**Percentage**
Mirage	19	18.6	Lighthouse	9	8.8
Chimera	13	12.7	Dolphin	8	7.8
Serenity	12	11.8	Elephant	8	7.8
Euphoria	11	10.8	Compass	7	6.7
Chaos	9	8.8	Mountain	6	5.9
Catalyst	8	7.8	Dream	6	5.9
Citadel	8	7.8	Microscope	6	5.9
Democracy	8	7.8	Feather	6	5.9
Abyss	8	7.8	Sundial	6	5.9
Justice	8	7.8	Rainbow	6	5.9

Furthermore, 117 responses by GPT4 (11.9%) across the 102 sets of words were a type of animal. In fact, only 10 sets of words out of the total of 102 (i.e., 9.8%) did *not* contain at least one animal in the set. A further 96 responses (10.1%) were geographic features or concepts (e.g., sandstorm, volcano, horizon). GPT3.5 showed a similar tendency to category repetition. This begs the question: if the essence of creativity is the production of novelty, then how can predictable results be classified as creative? If GPT4 is unreliable (in the sense that its responses have a large variance) and if it is predictable (in the sense that its responses follow a pattern), then can the hypothesis really be supported?

All of these considerations are tempered by the fact that divergent thinking, while necessary for creativity (and a very common way that creativity is operationalised, even in research literature), is nevertheless *not sufficient* (Guilford, 1950; Runco & Acar, 2012). Add to this the lack of *autonomy* in AI (Guckelsberger et al., 2017; Lamb et al., 2018) and we have, at best, a tool with a limited capacity to support one part of the broader creative problem-solving process.

Thus, it seems difficult to find unequivocal support for the hypothesis that *AI is more creative than humans*. ChatGPT, in the form of GPT4, has, on average, a higher verbal divergent production capability than most (85% of) humans, but this is highly variable, and somewhat predictable, making ChatGPT, at best, an unreliable source of verbal divergent production.

Conclusions

To explore the hypothesis that AI is more creative than humans, this study compared the verbal divergent thinking ability of ChatGPT (GPT3.5 and GPT4) with a large human norm, using the Divergent Association Task. The study found that, while both GPT3.5 and GPT4, on average, surpass the mean DAT scores of a human sample, this is tempered by issues of unreliability and predictability. In the strictest statistical sense, ChatGPT may perform better at verbal divergent production than humans, however, the nature of the performance of ChatGPT draws attention to the significant differences between “thinking divergently” and “being creative”. Future work should examine not only other large language models (i.e.,

Bard and Claude) but also possible new versions of ChatGPT. Nevertheless, the results reported in this paper suggest that the greatest promise of AI for creative tasks lies not as a replacement for human creativity, but as a support to human divergent thinking. Even then, tools such as ChatGPT need to be used carefully, with a clear understanding of their weaknesses and limitations. To paraphrase Mark Twain, the reports of the death of human creativity may be said to have been greatly exaggerated!

Lift Learning

Could AI replicate your creativity? Engage with Professor David Cropley as he discusses the rise of ChatGPT and its ability to "think" creatively at this article's companion LIFT Learning site. Professor Cropley explains the process of divergent thinking as a measure of creativity and outlines why your creative role is probably still safe for now. The LIFT Learning site is available at https://lift.c3l.ai/courses/course-v1:LEARNINGLETTERS+0113+2023

Funding

No funding was received for the conduct of this research.

Disclosure of conflicts of interest

The author reports no potential conflict of interest.

Disclosure of the use of AI-assisted technologies during writing

The author used ChatGPT (GPT4: Code Generator) for the purpose of creating Figure 2. The author takes full responsibility for the content.

About the author

David Cropley is the Professor of Engineering Innovation at the University of South Australia. He has been conducting research in creativity and innovation for more than 25 years. Among his interests are the role of creativity in engineering and design, the measurement of product creativity, the use of AI to automate the assessment of creativity, and the concept of deliberate misuse of creativity to cause harm (malevolent creativity). He is the author/co-author of 10 books on creativity including (with A. J. Cropley) *Core Capabilities for Industry 4.0 – Foundations of the Cyber-Psychology of Engineering*. Bielefeld, Germany: Wbv Media (2021). His latest research interrogates the many and varied claims about the creative abilities of large language models like ChatGPT.

ORCID: 0000-0002-7964-6538

References

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods, 53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Bloom, B. S. (1984). The two sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (2nd ed.). Pearson Education.

Cropley, D. H., & Marrone, R. L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000510

Cropley, D. H., Medeiros, K. E., & Damadzic, A. (2022). Creativity and artificial intelligence: The integration of human and artificial cognition. In D. Henriksen & P. Mishra (Eds.), *Creative provocations: Speculations on the future of creativity, technology, and learning* (pp. 19–34). Springer.

Cropley, D. H., Patston, T., Marrone, R. L., & Kaufman, J. C. (2019). Essential, unexceptional and universal: Teacher implicit beliefs of creativity. *Thinking Skills and Creativity, 34*, Article 100604. https://doi.org/10.1016/j.tsc.2019.100604

Du Sautoy, M. (2019). *The creativity code*. Harvard University Press.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gioti, A.-M. (2021). Artificial intelligence for music composition. In E. R. Miranda (Ed.), *Handbook of artificial intelligence for music: Foundations, advanced approaches, and developments for creativity* (pp. 53–73). Springer. https://doi.org/10.1007/978-3-030-72116-9_3

Guckelsberger, C., Salge, C., & Colton, S. (2017, June 19-23). *Addressing the "why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency.* [Conference presentation] International Conference on Computational Creativity, Atlanta, GA.

Guilford, J. P. (1950). Creativity. *American Psychologist, 5*, 444–454.

Henriksen, D., Woo, L. J., & Mishra, P. (2023). Creative uses of ChatGPT for education: A conversation with Ethan Mollick. *TechTrends*, *67*, 595–600.

Kirkpatrick, K. (2023). Can AI demonstrate creativity? *Communications of the ACM, 66*(2), 21–23. https://doi.org/10.1145/3575665

Lamb, C., Brown, D. G., & Clarke, C. L. (2018). Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR), 51*(2), Article 28, 1–34. https://doi.org/10.1145/3167476

Marrone, R., Cropley, D. H., & Wang, Z. (2022). Automatic assessment of mathematical creativity using natural language processing. *Creativity Research Journal*, 1–16. https://doi.org/10.1080/10400419.2022.2131209

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences, 118*(25). https://doi.org/10.1073/pnas.2022340118

Plucker, J., Makel, M., & Qian, M. (2019). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (2nd ed., pp. 44–68). Cambridge University Press.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., & Acar, S. (2019). Divergent thinking. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (2nd ed., pp. 224–254). Cambridge University Press.

Succi, C., & Canovi, M. (2020). Soft skills to enhance graduate employability: Comparing students and employers' perceptions. *Studies in Higher Education, 45*(9), 1834–1847. https://doi.org/10.1080/03075079.2019.1585420

Tabachnik, B. G., & Fidell, S. L. (2013). *Using multivariate statistics*. Pearson Education.

Torrance, E. P. (1999). *Torrance test of creative thinking: Norms and technical manual*. Scholastic Testing Services.

Urban, K. K., & Jellen, H. G. (1996). *Test for Creative Thinking - Drawing Production (TCT-DP)*. Swets and Zeitlinger.