Creativity Assessment in the Age of AI

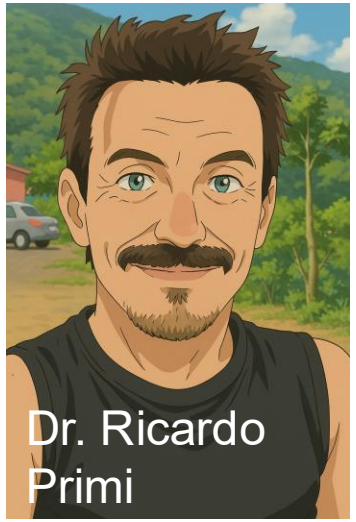
Psychometric and Computational Approaches

Ricardo Primi

University of São Francisco & Edulab21, Ayrton Senna Institute, Brazil

Creativity Assessment in the Age of AI: Psychometric and Computational Approaches

- This short course introduces cutting-edge methods in creativity assessment, integrating psychometrics and artificial intelligence. We begin by addressing rater design challenges, focusing on Balanced Incomplete Block Designs (BIBD) to optimize rating efficiency. Next, we explore the use of the Many-Facet Rasch Model (MFRM) to model and adjust for rater severity, bias, and consistency in scoring. The course then turns to automated scoring, showing how artificial intelligence – particularly Large Language Models (LLMs) – can be used to evaluate creative responses. We also demonstrate how to model and compare human and AI-based ratings using MFRM. Throughout the course, we will share and explain R code for implementing each of these methods, offering participants hands-on tools for scalable, valid, and reliable creativity assessment.



Creativity Assessment in the Age of AI: Psychometric and Computational Approaches

Campinas

About the Presenter

Associate Professor, Graduate Program in Psychological Assessment, University of São Francisco, Brazil

Lead Researcher, eduLab21 – Ayrton Senna Institute's Science Lab for Education (since 2014)

Member, OECD Technical Advisory Group on Socio-Emotional Skills

Member, PISA 2021 Questionnaire Expert Group (QEG), coordinated by ETS and OECD

Visiting Scholar, UC Berkeley (Institute of Personality and Social Research, CAPES Fellow)

Visiting Research Scholar, Stanford University (Lemann Center for Educational Innovation (2020)

Recent research focuses on AI applications in assessing personality and creativity

Summary

- Theory of Creativity (Creativity 101)
  - Basic concepts on creativity
- Assessment of creativity
  - Types of measures
  - Challenges of the assessment
- Psychometrics of divergent thinking tasks
  - Test of Metaphor Creation
  - Many Facet Rasch Models
  - Efficient Rating Designs
- Coding time
  - BIBD
  - MFRM
- Basic concepts about IA and Transformers
- Use of Artificial Intelligence in Creativity Assessment

Creativity 101

What is creativity?

- “Creativity is the interaction among aptitude, process, and environment by which an individual or group produces a perceptible product that is both **novel** and **useful** as defined within a social context” (Plucker et al., 2004, p. 90).

What is creativity?

- In PISA 2022 is defined as

  the competence to engage productively in the generation, evaluation and improvement of ideas, that can result in original and effective solutions, advances in knowledge and impactful expressions of imagination.

- **What Makes Creativity Assessment Unique?**
- Focus on **divergent thinking**: tasks invite multiple valid responses, not a single correct answer
- Scoring emphasizes:
  - **Appropriateness** – relevance and alignment with the task
  - **Originality** – how rare or novel the response is
  - **Effectiveness/Value** – the usefulness or impact of the idea

Creative potential vs Creative Achievement

- **Creative potential** is an individual's capacity to generate novel and potentially useful ideas, solutions, or artistic expressions.
- **Creative achievement** refers to the concrete outcomes or products resulting from creative potential, which are recognized as novel and valuable by others, or have a demonstrable impact (Benedek, 2024, Jauk et al., 2014).

Creative potential vs Creative Achievement

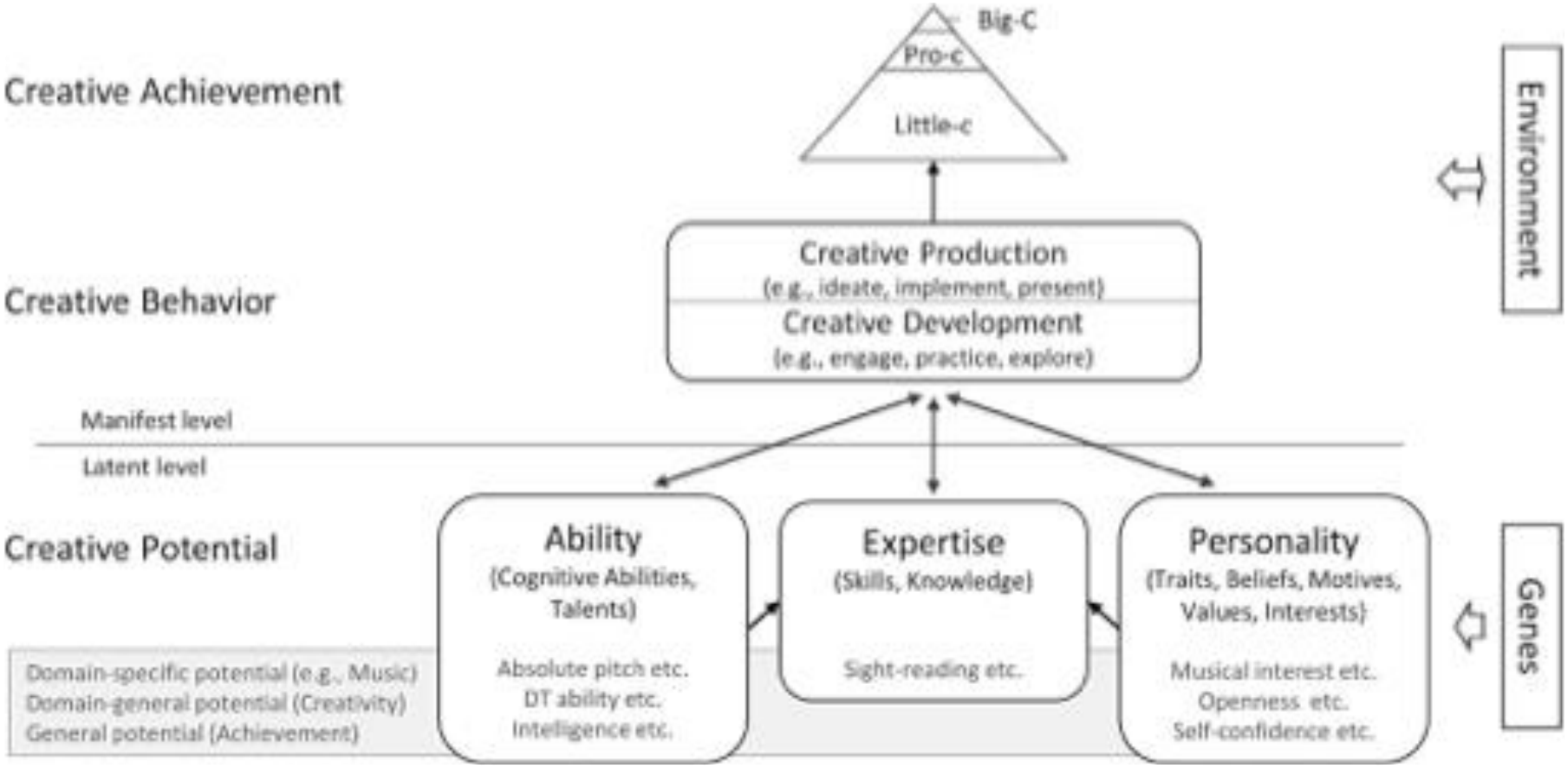


Fig. 2. A componential model of the creative potential-achievement relationship.

Creative achievement distinctions (Kaufman & Beghetto, 2009, 4-C model)

Mini-c creativity is defined as the novel and personally meaningful interpretation of experiences, actions, and events. It creative process as it relates to personal growth and learning rather than on the creative product itself.

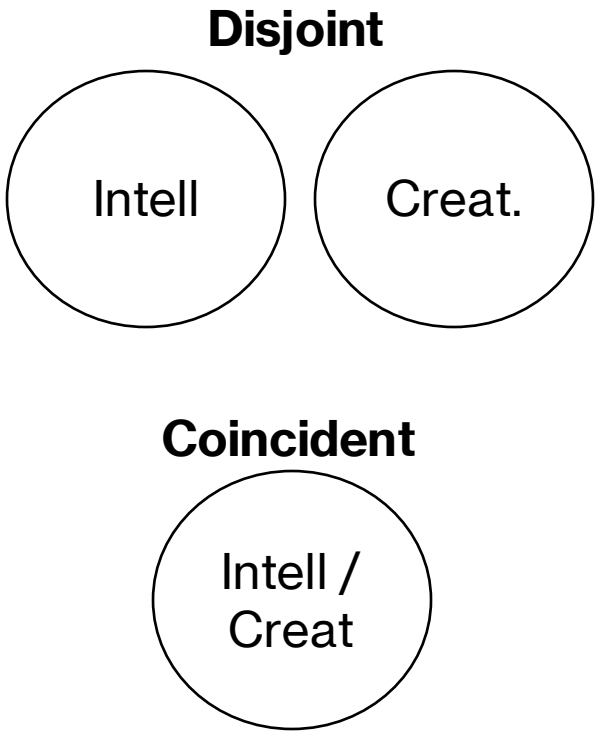
Little-c creativity refers to everyday creative expressions and problem-solving that have value to the individual and sometimes to others.

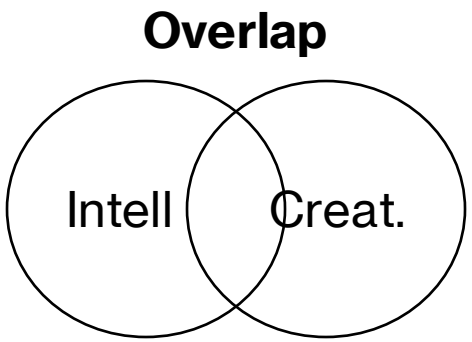
Pro-c creativity describes creativity at a professional or expert level. Individuals at this level have developed their skills through deliberate practice and training and produce creative work recognized by peers and professionals

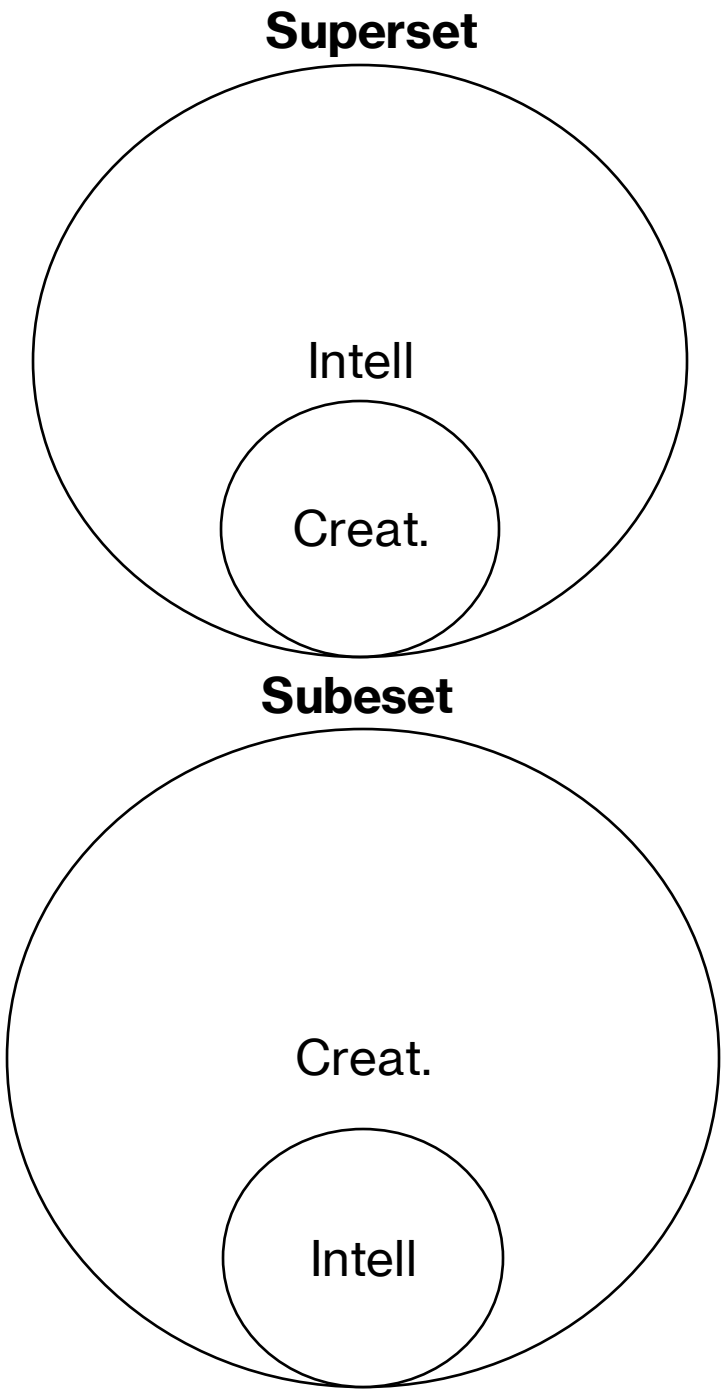
Big-C creativity is reserved for eminent creative achievements that have a significant, lasting impact on a domain or society, such as groundbreaking scientific discoveries, artistic masterpieces, or major innovations.

Creativity and Intelligence

Sternberg and O'Hara (1998)







Threshold Hypothesis for divergent thinking

Intelligence 41 (2013) 212–221

Contents lists available at SciVerse ScienceDirect

Intelligence

journal homepage:

The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection

Emanuel Jauk *, Mathias Benedek, Beate Dunst, Aljoscha C. Neubauer

Department of Psychology, University of Graz, Austria

ARTICLE INFO

Article history:
Received 1 August 2012
Received in revised form 12 March 2013
Accepted 14 March 2013
Available online xxxx

Keywords:
Threshold hypothesis
Intelligence
Creativity
Segmented regression
Breakpoint detection

ABSTRACT

The relationship between intelligence and creativity has been subject to empirical research for decades. Nevertheless, there is yet no consensus on how these constructs are related. One of the most prominent notions concerning the interplay between intelligence and creativity is the threshold hypothesis, which assumes that above-average intelligence represents a necessary condition for high-level creativity. While earlier research mostly supported the threshold hypothesis, it has come under fire in recent investigations. The threshold hypothesis is commonly investigated by splitting a sample at a given threshold (e.g., at 120 IQ points) and estimating separate correlations for lower and upper IQ ranges. However, there is no compelling reason why the threshold should be fixed at an IQ of 120, and to date, no attempts have been made to detect the threshold empirically. Therefore, this study examined the relationship between intelligence and different indicators of creative potential and of creative achievement by means of segmented regression analysis in a sample of 297 participants. Segmented regression allows for the detection of a threshold in continuous data by means of iterative computational algorithms. We found thresholds only for measures of creative potential but not for creative achievement. For the former the thresholds varied as a function of criteria. When investigating a liberal criterion of ideational originality (i.e., two original ideas), a threshold was detected at around 100 IQ points. In contrast, a threshold of 120 IQ points emerged when the criterion was more demanding (i.e., many original ideas). Moreover, an IQ of around 85 IQ points was found to form the threshold for a purely quantitative measure of creative potential (i.e., ideational fluency). These results confirm the threshold hypothesis for qualitative indicators of creative potential and may explain some of the observed discrepancies in previous research. In addition, we obtained evidence that once the intelligence threshold is met, personality factors become more predictive for creativity. On the contrary, no threshold was found for creative achievement, i.e. creative achievement benefits from higher intelligence even at fairly high levels of intellectual ability.

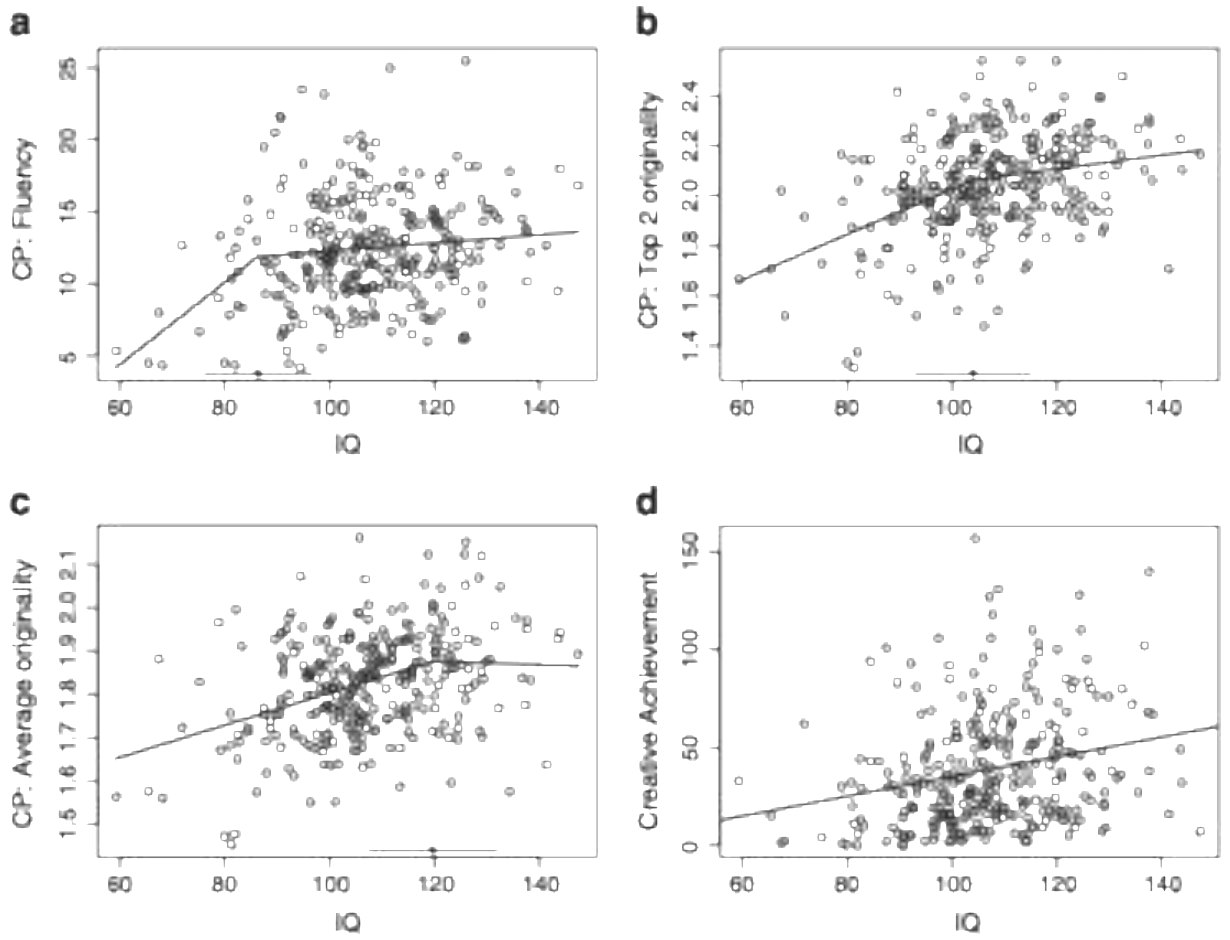


Fig. 1. Breakpoint models for the fluency score (a), the Top 2 originality score (b), and the average originality score (c). Linear model for creative achievement (d). Horizontal lines indicate 95% CI of the breakpoint. CP: creative potential.

Threshold Hypothesis for divergent thinking

European Journal of Education and Psychology
2010, Vol. 3, Nº 1 (Págs. 103-115)

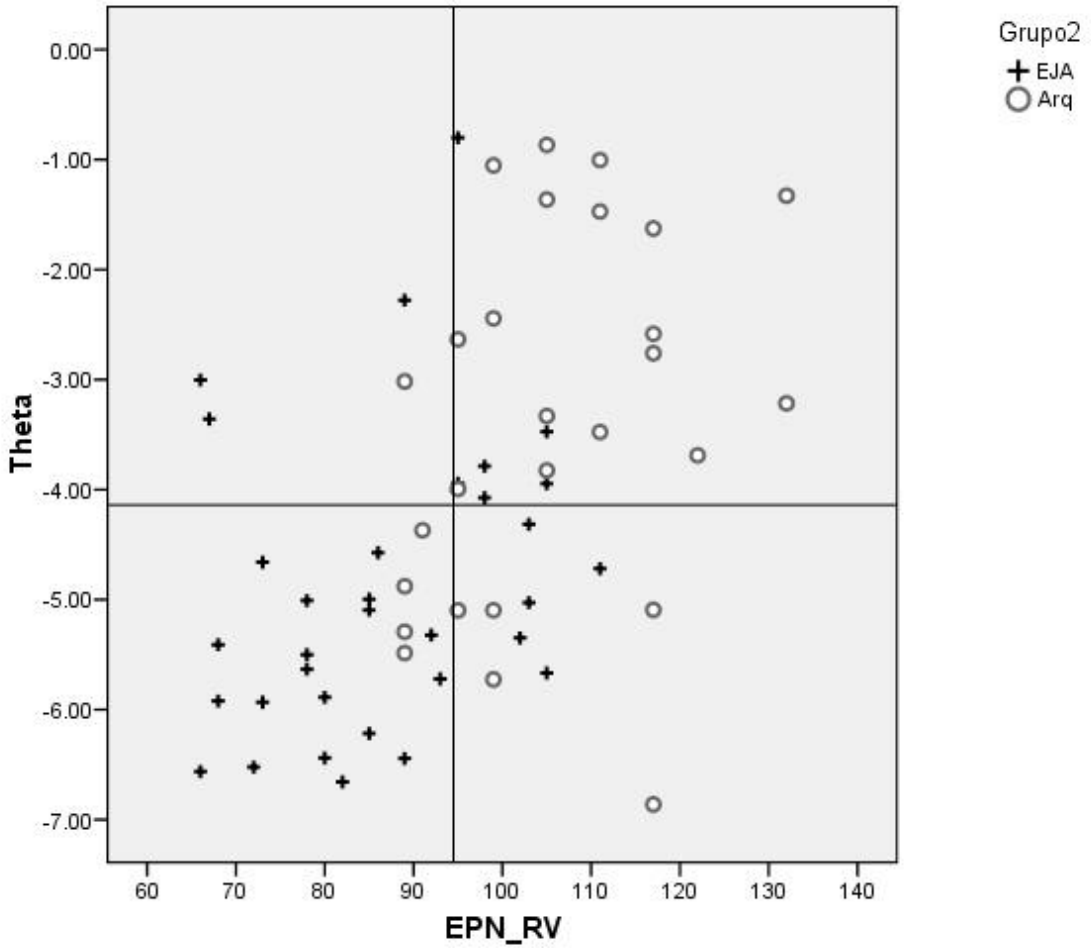
ISSN 1888-8992 // www.ejep.es

Metaphor Creation: A Measure of Creativity or Intelligence?

Débora Pereira de Barros[1], Ricardo Primi[1], Fabiano Koich Miguel[2],
Leandro S. Almeida[2] and Ema P. Oliveira[3]
[1]Universidade São Francisco (Brasil), [2]Universidade do Minho (Portugal),
[3]Universidade da Beira Interior (Portugal)

The goal of the present study was to verify whether the "Metaphor Creation Test" would really be a measure with unique characteristics of creativity, or a different way of evaluating constructs already known as intelligence. Two differentiated groups were considered: group 1 was comprised by 90 late course students, and group 2 included 73 undergraduate students from Architecture and Urbanism courses. The results showed lower correlation between metaphors production and abstract reasoning (r= .31) comparing with verbal reasoning test (r= .48). Correlations between the constructs reproduced what was already found in other studies, that is, intelligence and creativity are related, but not strongly enough to affirm that they are the same construct; therefore they are different but related constructs.

Key words: Creativity assessment, creativity and intelligence, psychological assessment, metaphor creation.



Necessary Condition (NCA) Hypothesis for divergent thinking

Contents lists available at ScienceDirect

Intelligence

Is creativity without intelligence possible? A Necessary Condition Analysis

Maciej Karwowski [a,*], Jan Dul [b], Jacek Gralewski [a], Emanuel Jauk [c], Dorota M. Jankowska [a], Aleksandra Gajda [a], Michael H. Chruszczewski [d], Mathias Benedek [c]

[a] The Maria Grzegorzewska University, Warsaw, Poland
[b] Erasmus University, The Netherlands
[c] University of Graz, Austria
[d] University of Warsaw, Poland

ARTICLE INFO

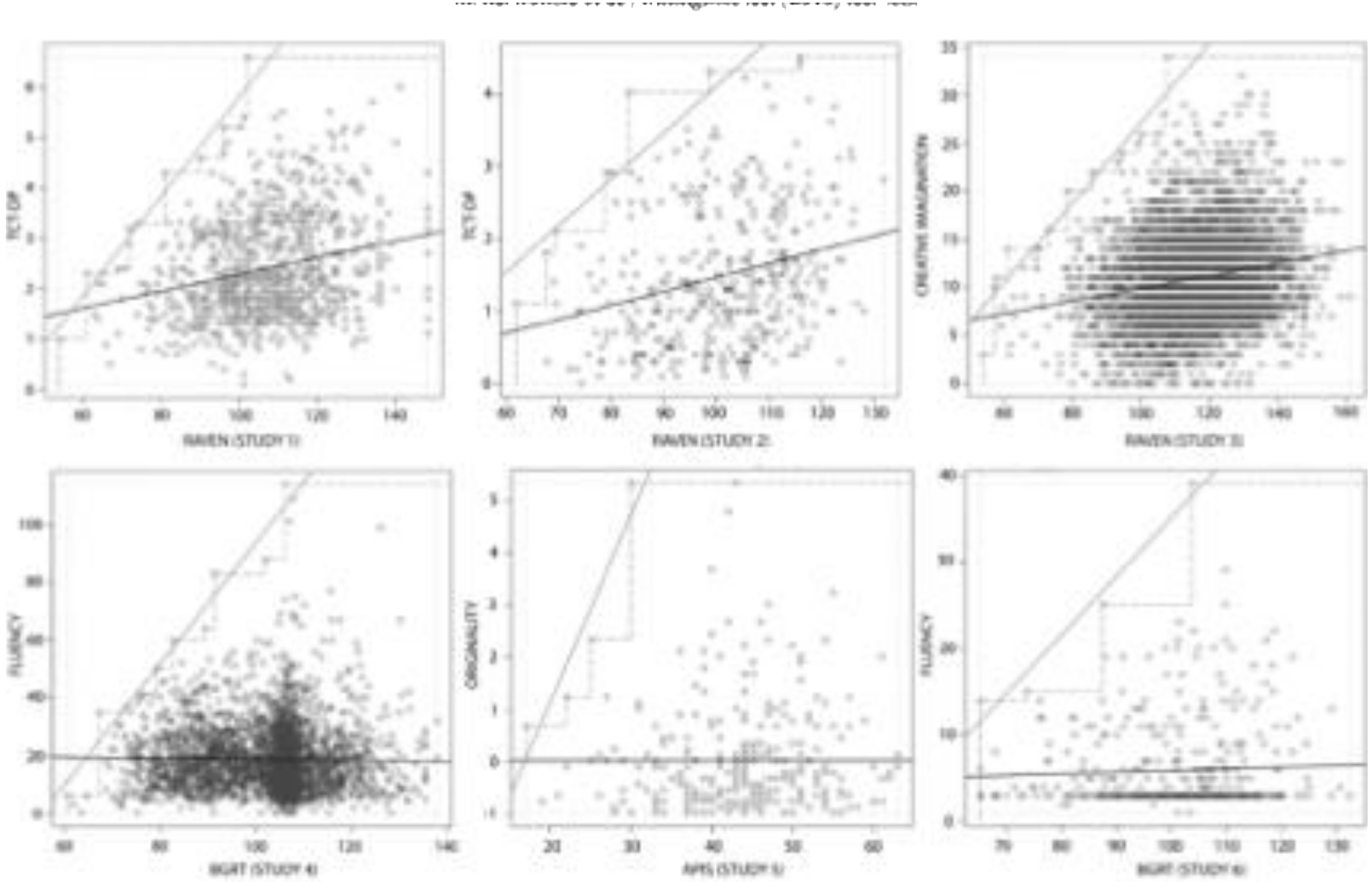
Article history:
Received 22 January 2016
Received in revised form 7 April 2016
Accepted 26 April 2016
Available online xxxx

Keywords:
Intelligence
Creativity
Threshold hypothesis
Necessary condition hypothesis
Necessary Condition Analysis

ABSTRACT

This article extends the previous studies on the relationship between intelligence and creativity by providing a new methodology and an empirical test of the hypothesis that intelligence is a necessary condition for creativity. Unlike the classic threshold hypothesis, which assumes the existence of a curvilinear relationship between intelligence and creativity, the Necessary Condition Analysis (Dul, 2016) focuses on and quantifies the overall shape of the relationship between intelligence and creativity. In eight studies (total $N = 12,255$), using different measures of intelligence and creativity, we observed a consistent pattern that supports the necessary-but-not-sufficient relationship between these two constructs. We conclude that although evidence concerning the threshold hypothesis on the creativity–intelligence relationship is mixed, the "necessary condition hypothesis" is clearly corroborated by the results of appropriate tests.

Necessary Condition (NCA) Hypothesis for divergent thinking



Necessary Condition (NCA) Hypothesis for divergent thinking

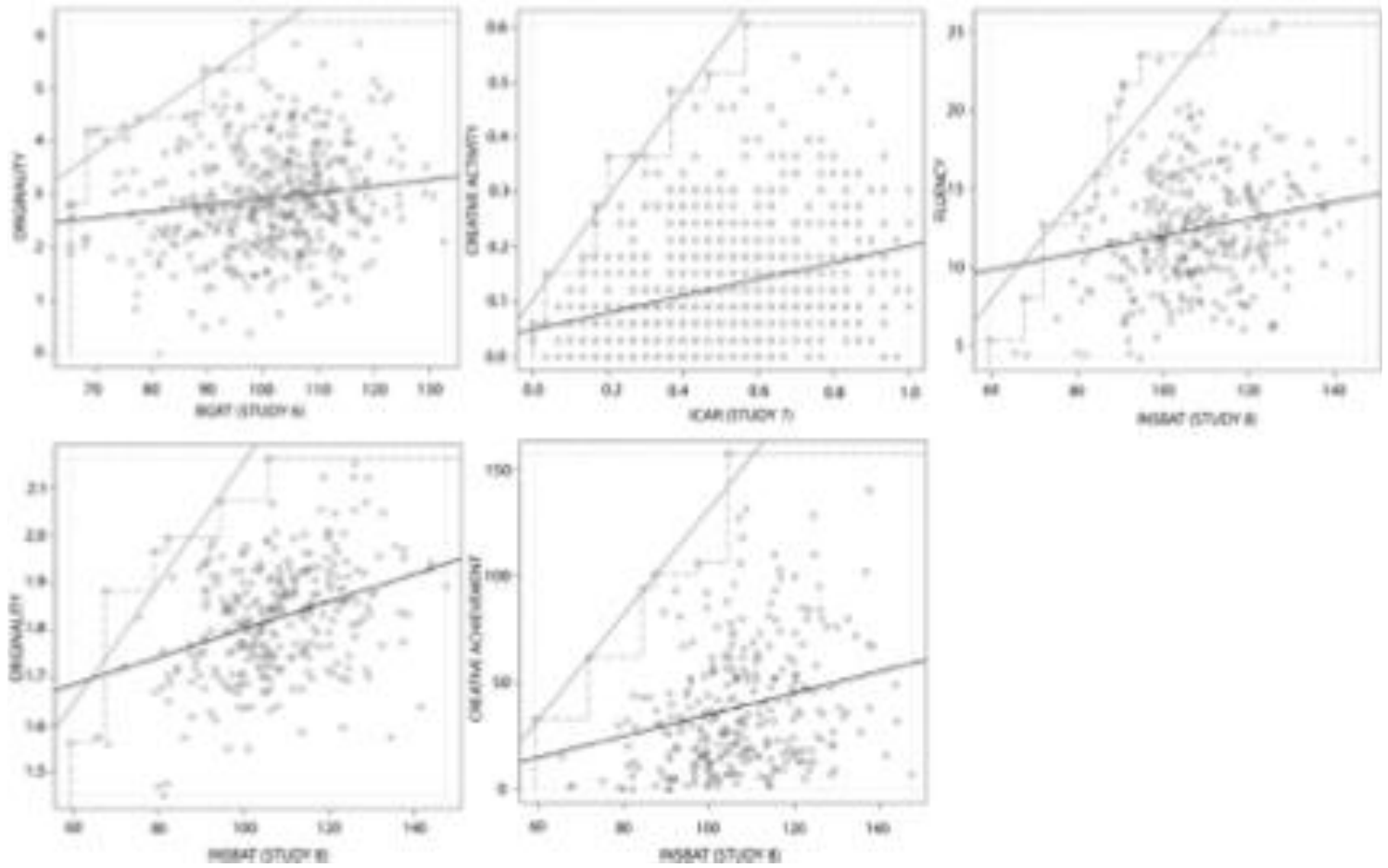
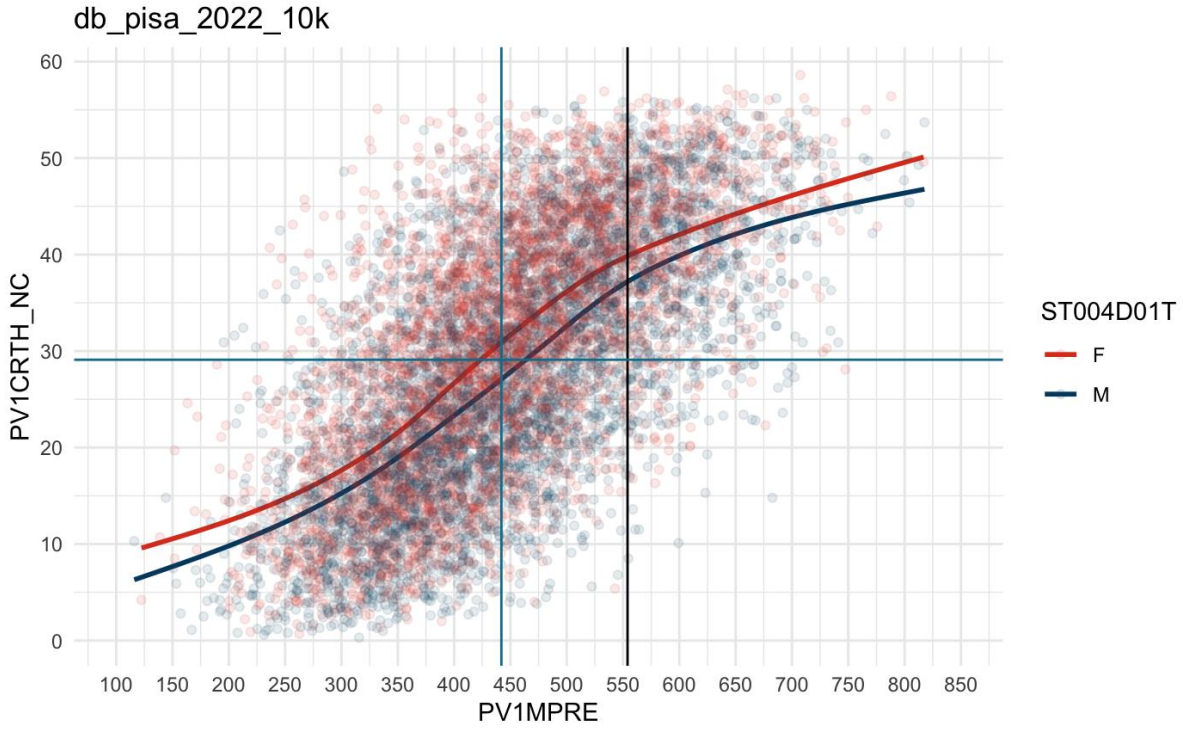
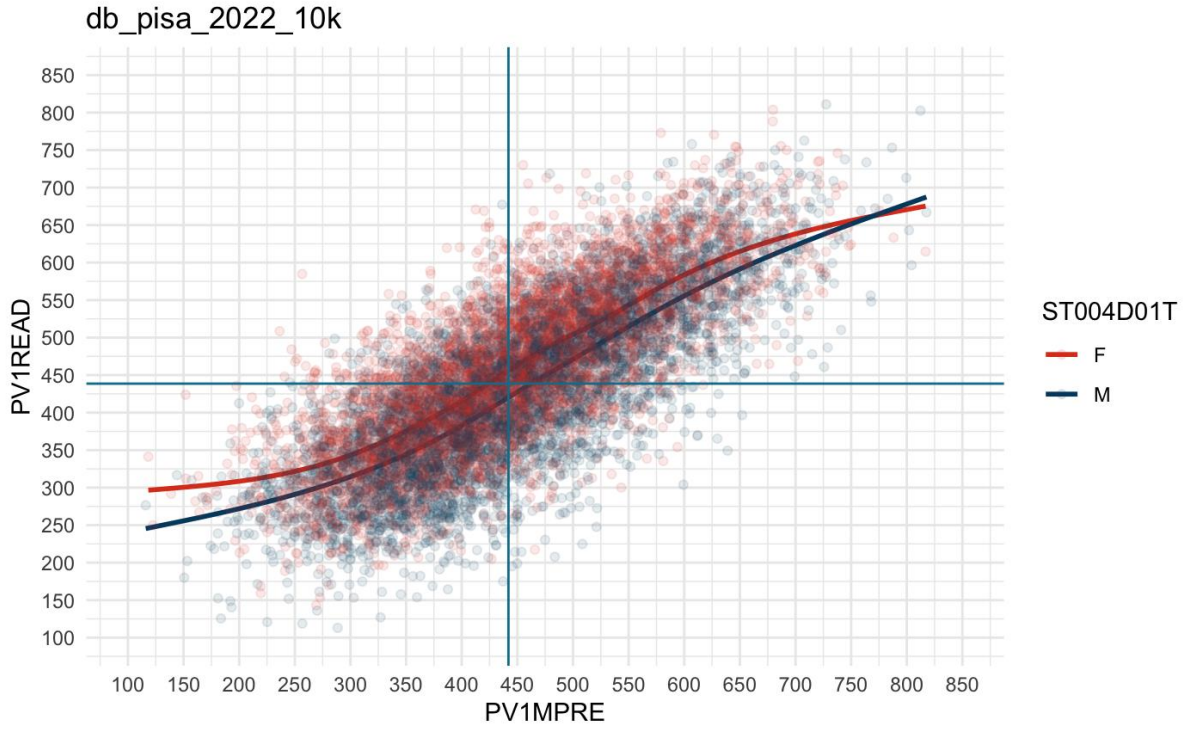


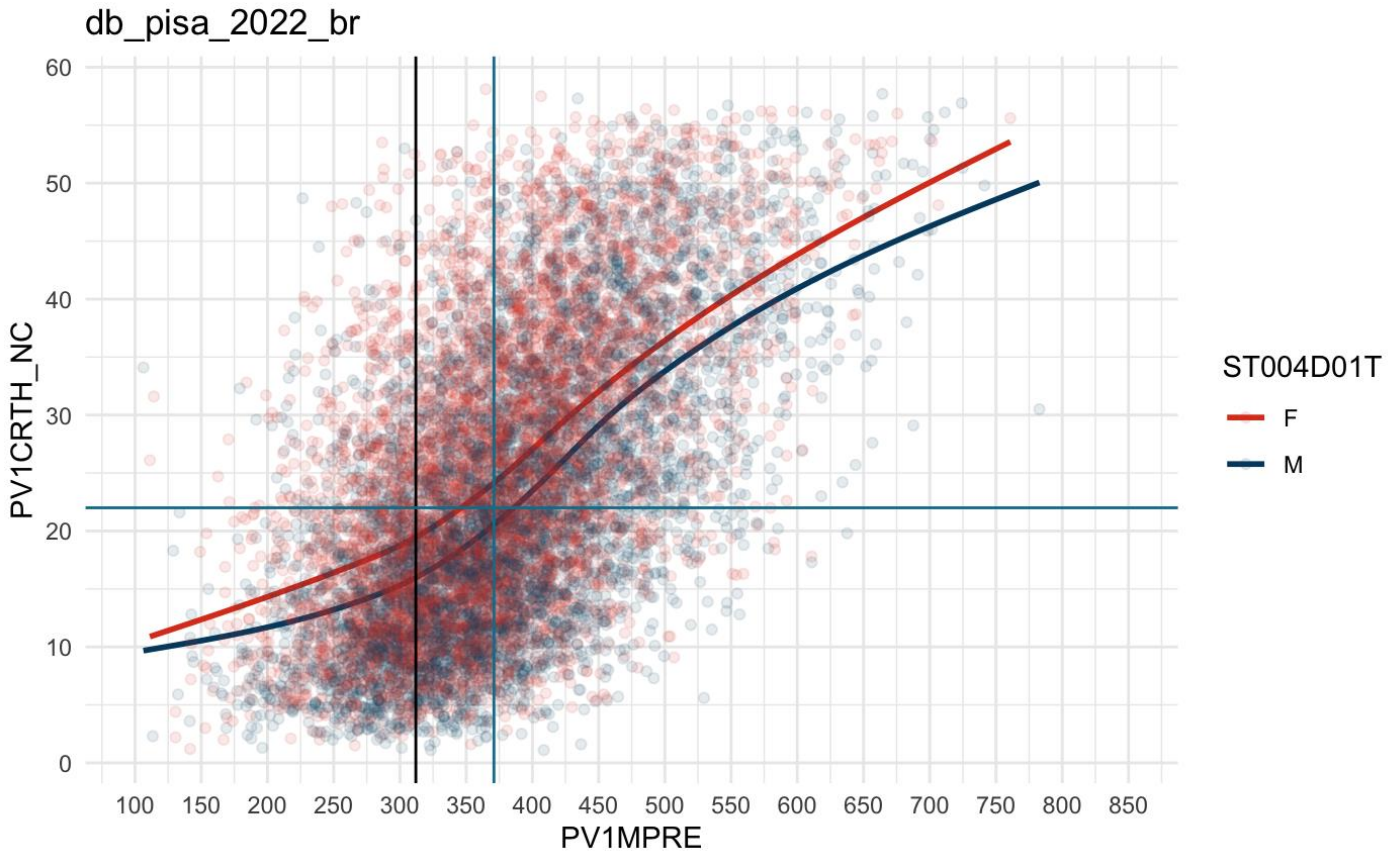
Fig. 2. Visualization of the NCA across eight studies and eleven scores. Black lines denote the linear (correlational) function, solid gray lines denote the CR-FDH ceiling line, and broken gray lines denote the CE-FDH ceiling line.

Threshold Hypothesis in PISA



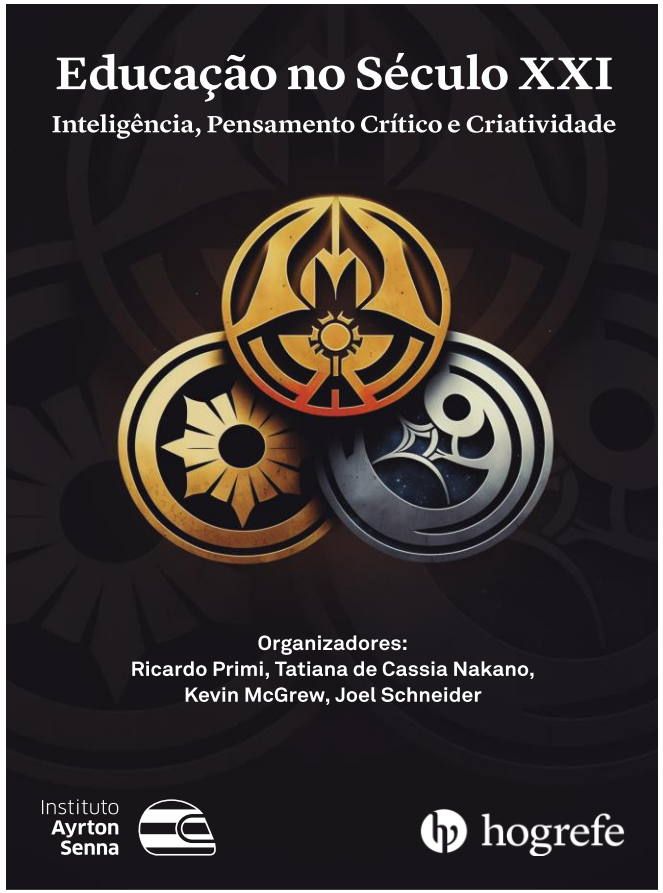


Threshold Hypothesis in PISA



Creativity in modern theory of Intelligence

Cattel-Horn-Carroll (CHC) model of intelligence



Contents lists available at ScienceDirect

Intelligence

Editorial

CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research

Kevin S. McGrew*

Woodcock-Muñoz Foundation, University of Minnesota, United States

ARTICLE INFO

Article history:
Received 4 August 2008
Received in revised form 8 August 2008
Accepted 9 August 2008
Available online 26 September 2008

Keywords:
Intelligence
Cognitive abilities
Factor analysis
John Horn
John Carroll
HCA
Gf–Gc theory
Cattell–Horn–Carroll Theory
CHC theory

ABSTRACT

During the past decade the Cattell–Horn Gf–Gc and Carroll Three-Stratum models have emerged as the consensus psychometric-based models for understanding the structure of human intelligence. Although the two models differ in a number of ways, the strong correspondence between the two models has resulted in the increased use of a broad umbrella term for a synthesis of the two models (Cattell–Horn–Carroll theory of cognitive abilities—CHC theory).

The purpose of this editorial is three-fold. First, I will describe the CHC framework and recommend that intelligence researchers begin using the CHC taxonomy as a common nomenclature for describing research findings and a theoretical framework from which to test hypotheses regarding various aspects of human cognitive abilities. Second, I argue that the emergence of the CHC framework should not be viewed as the capstone to the psychometric era of factor analytic research. Rather, I recommend the CHC framework serve as the stepping stone to reinvigorate the investigation of the structure of human intelligence.

Finally, the Woodcock-Muñoz Foundation Human Cognitive Abilities (HCA) project, which is an evolving, free, on-line electronic archive of the majority of datasets analyzed in Carroll's (1993) seminal treatise on factor analysis of human cognitive abilities, is introduced and described. Intelligence scholars are urged to access the Carroll HCA datasets to test and evaluate structural models of human intelligence with contemporary methods (confirmatory factor analysis). In addition, suggestions are offered for linking the analysis of contemporary data sets with the seminal work of Carroll. The emergence of a consensus CHC taxonomy and access to the original datasets analyzed by Carroll provides an unprecedented opportunity to extend and refine our understanding of human intelligence.

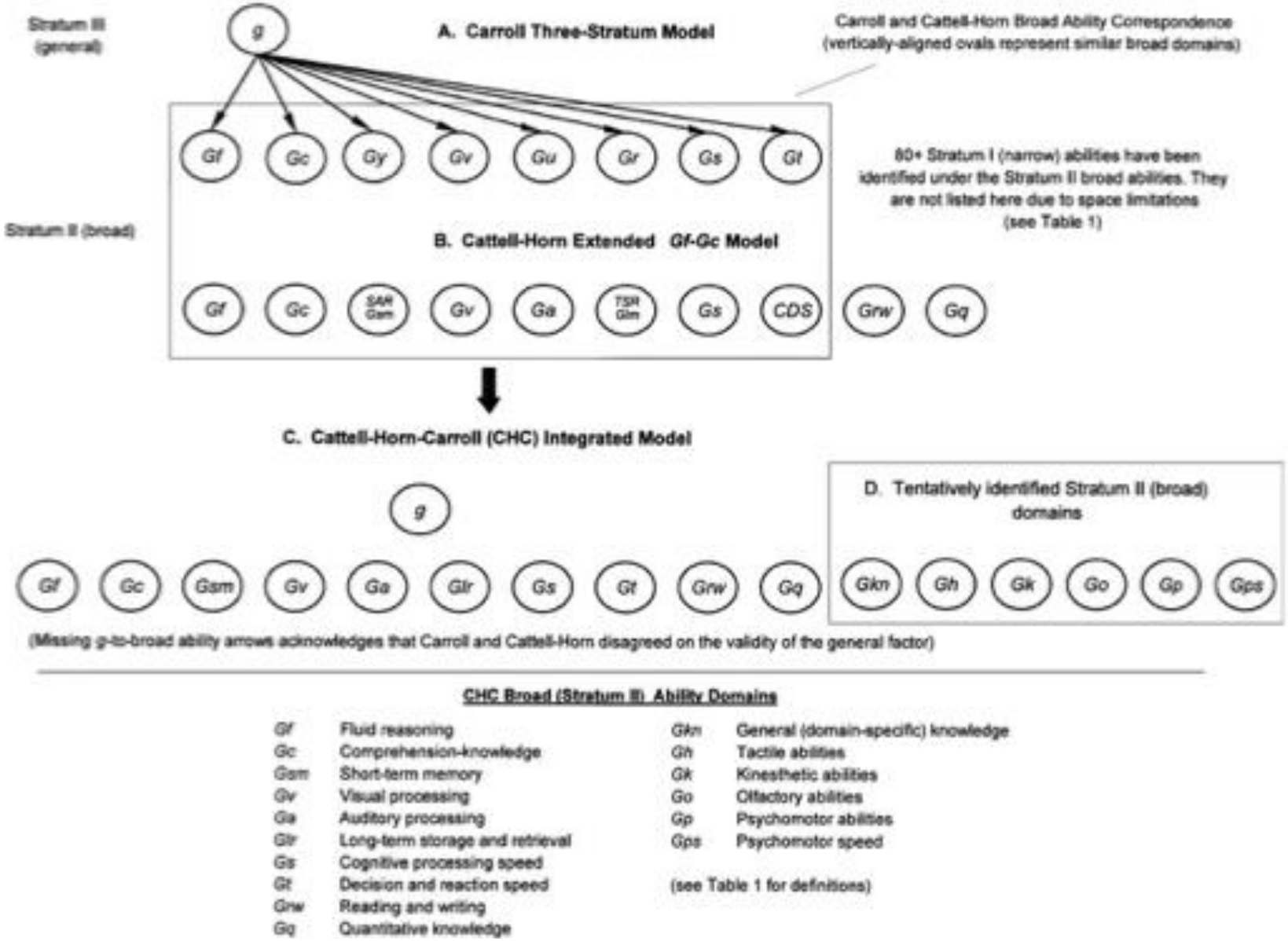


Fig. 1. Schematic representation and comparisons of Carroll's Three-Stratum, Cattell–Horn's Extended Gf–Gc, and the integrated Cattell–Horn–Carroll models of human cognitive abilities.

CHC Model

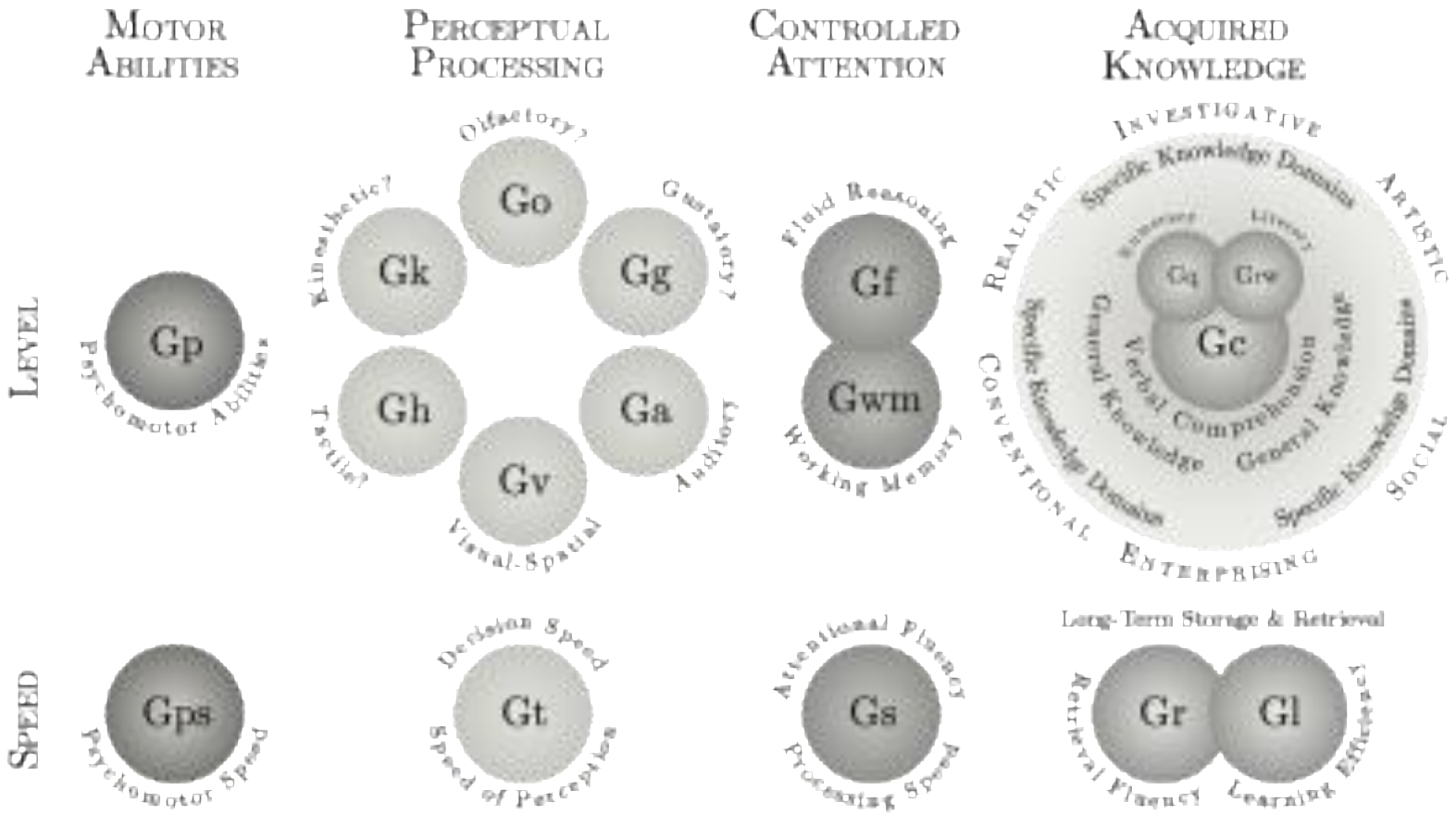
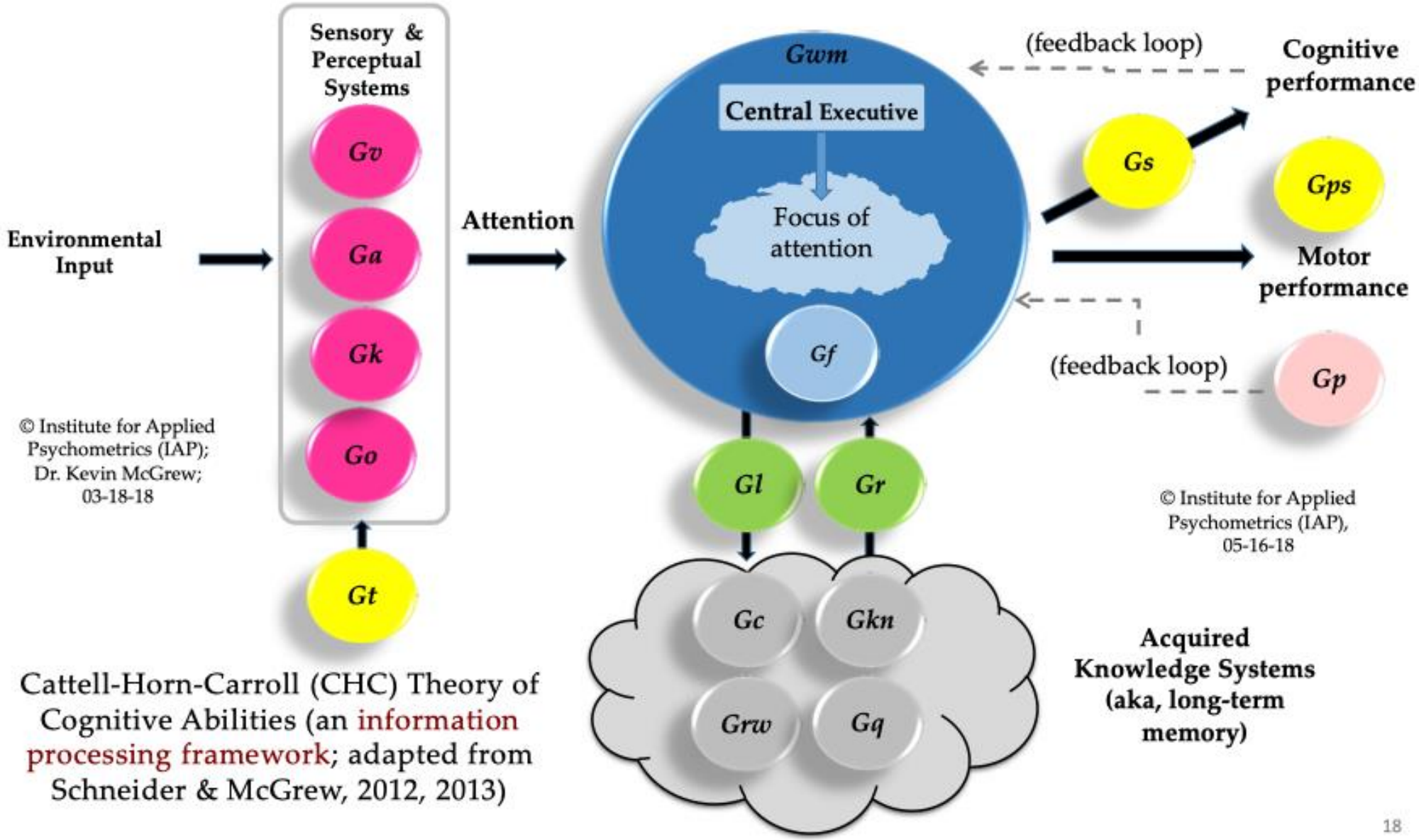
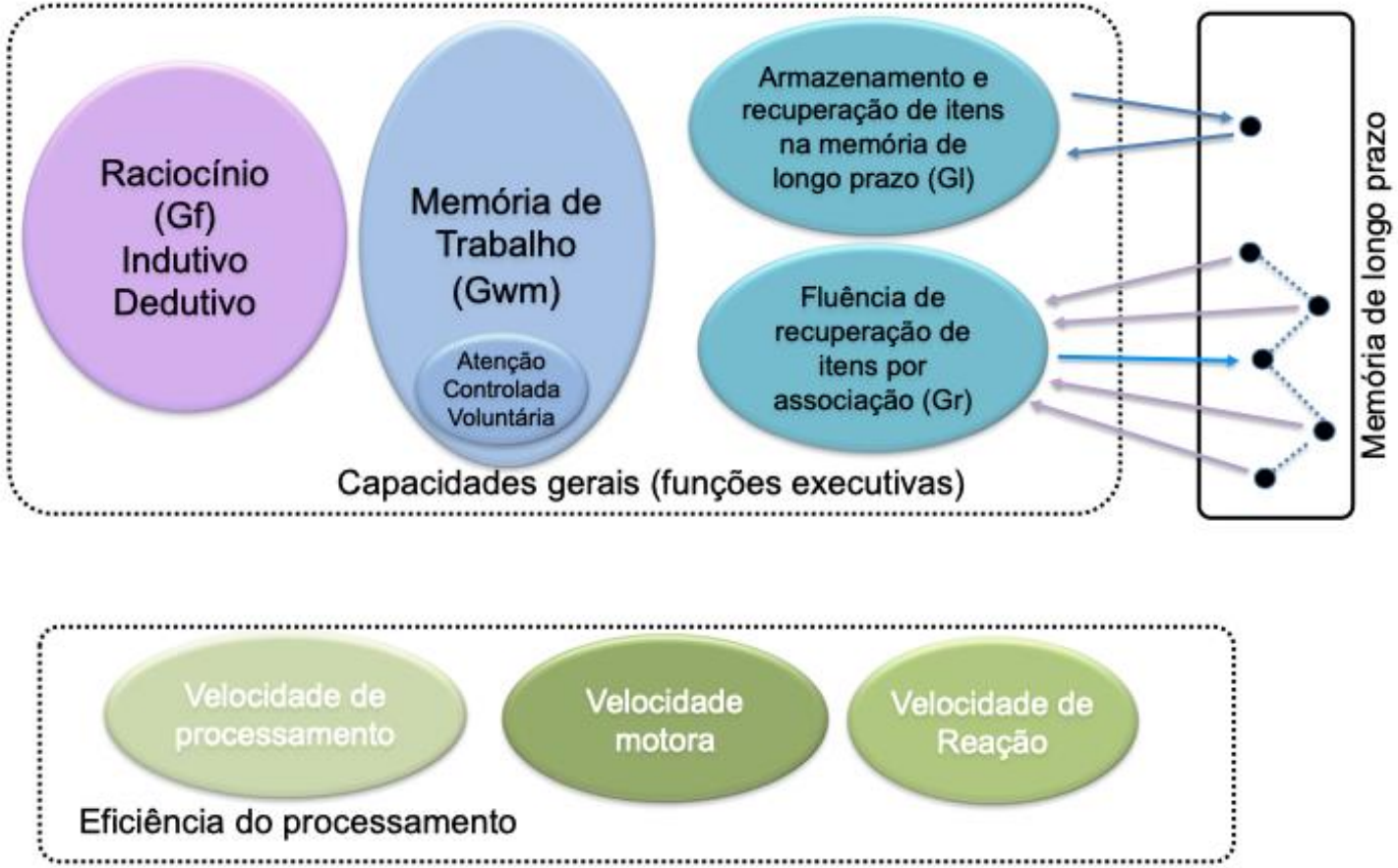


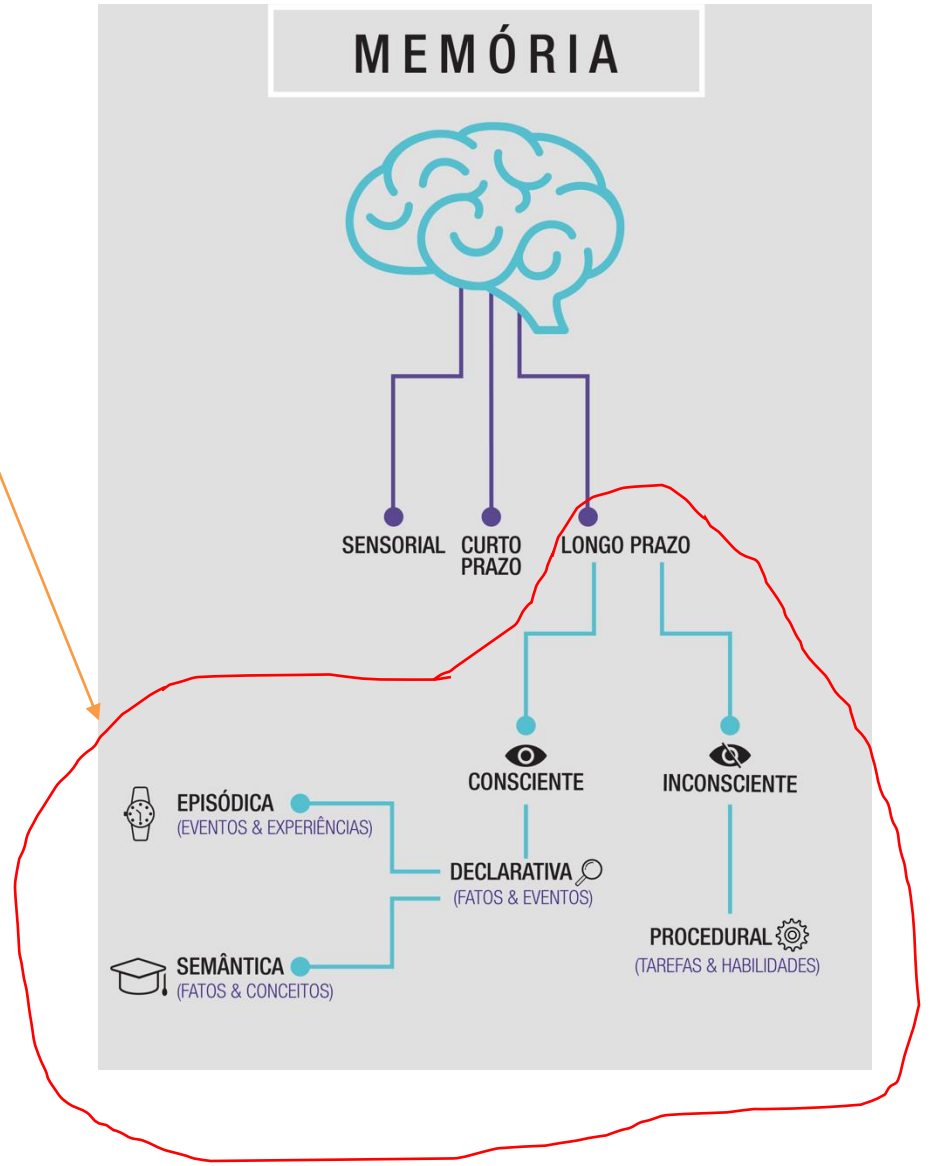
FIGURE 3.2. Conceptual groupings of CHC broad abilities.



Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities (an information processing framework; adapted from Schneider & McGrew, 2012, 2013)

Inteligência como processos cognitivos





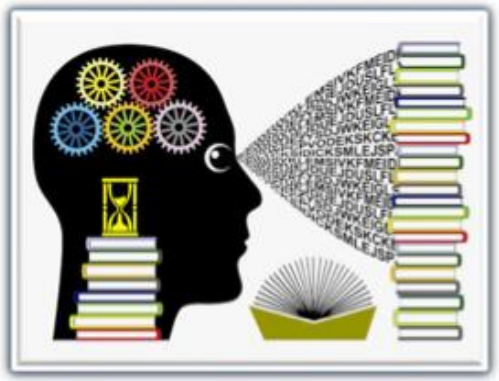
http://www.iapsych.com/glgr062116.pdf

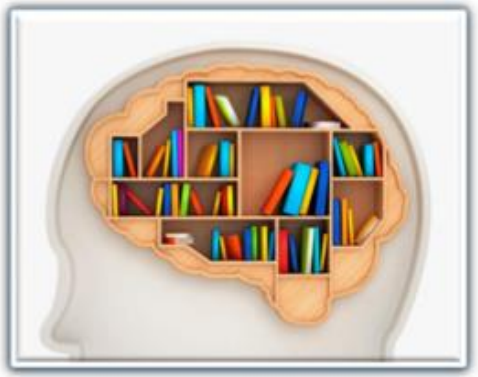
https://themindhub.com/research-reports

Gr

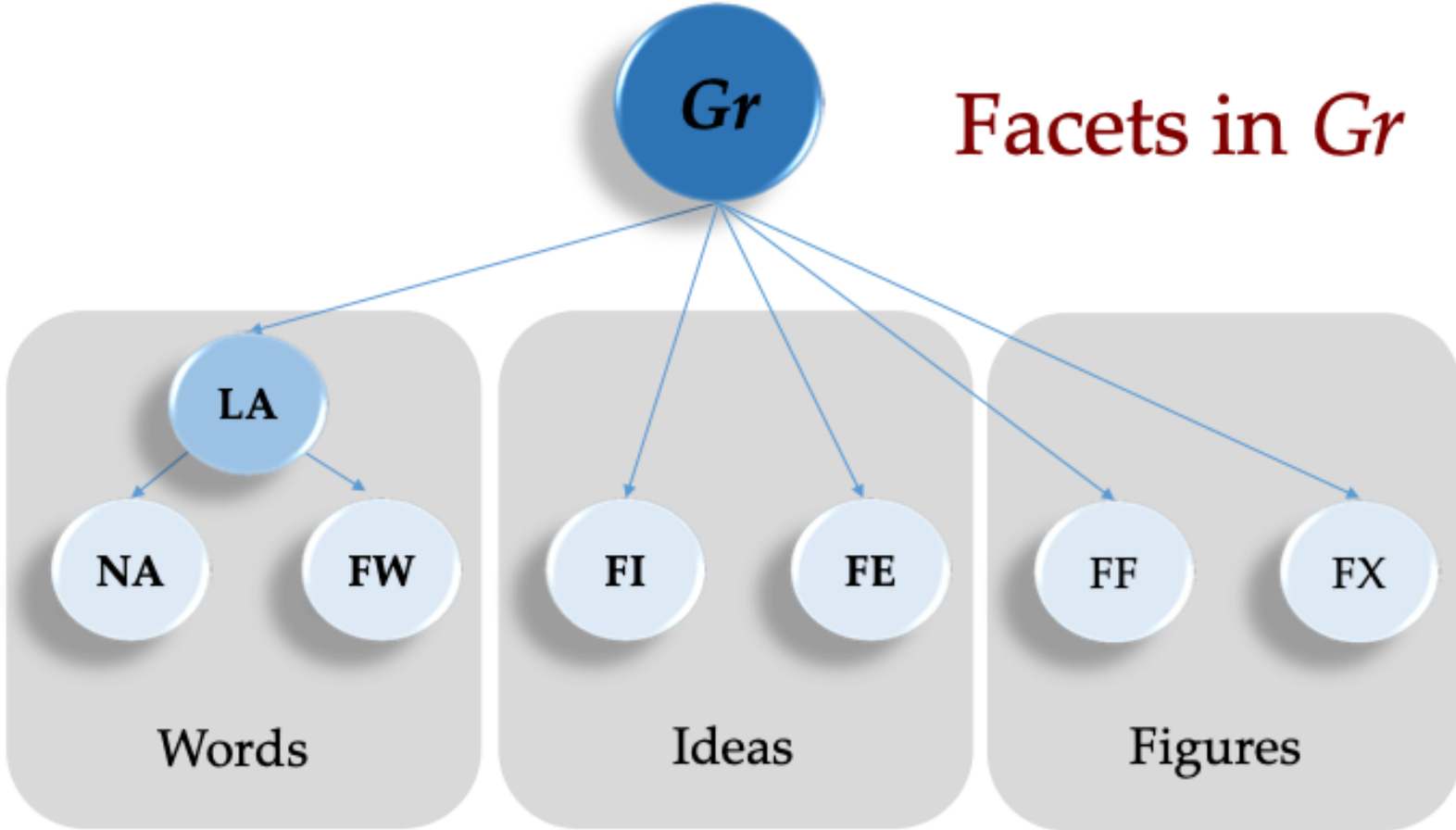
Retrieval fluency

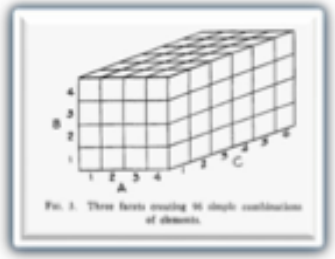
The rate and fluency at which individuals can access information stored in long-term memory.



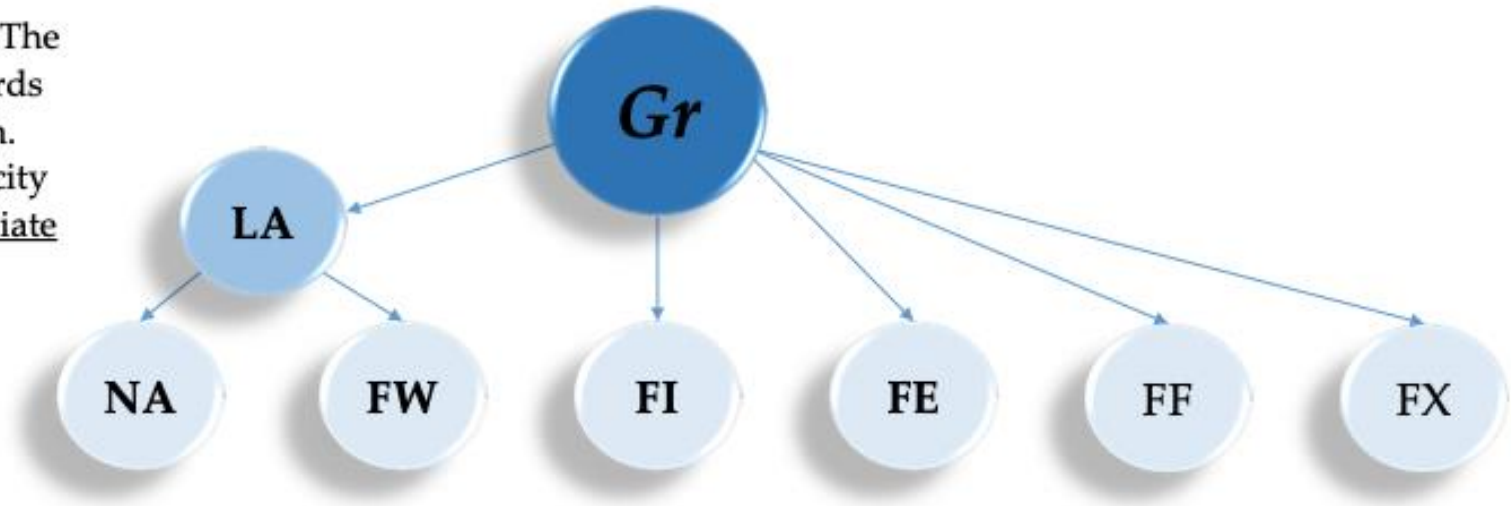


Facets in *Gr*





Speed of lexical access (LA): The ability to rapidly retrieve words from an individual's lexicon. Verbal efficiency or automaticity of lexical access. An intermediate stratum level ability.

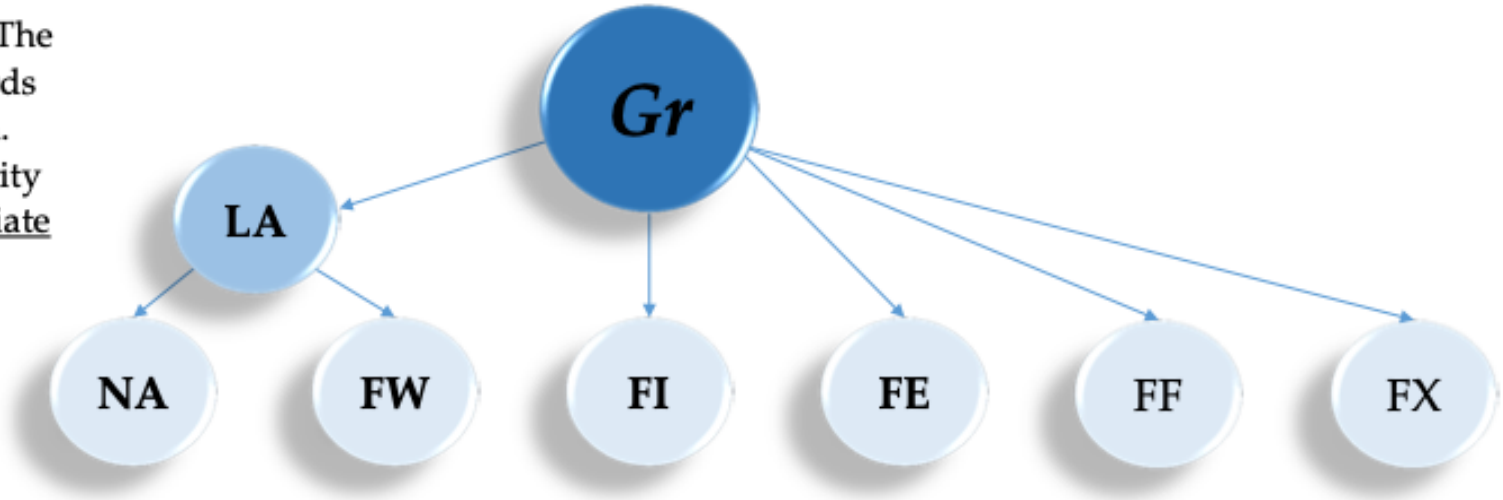


Naming facility (NA): The ability to rapidly call objects by their names.

Word fluency (FW): The ability to rapidly produce words that share a phonological (e.g., fluency of retrieval of words via a phonological cue) or semantic feature (e.g., fluency of retrieval of words via a meaning-based representation).

Ideational fluency (FI): The ability to rapidly produce a series of ideas, words, or phrases related to a specific condition or object.

Speed of lexical access (LA): The ability to rapidly retrieve words from an individual's lexicon. Verbal efficiency or automaticity of lexical access. An intermediate stratum level ability.



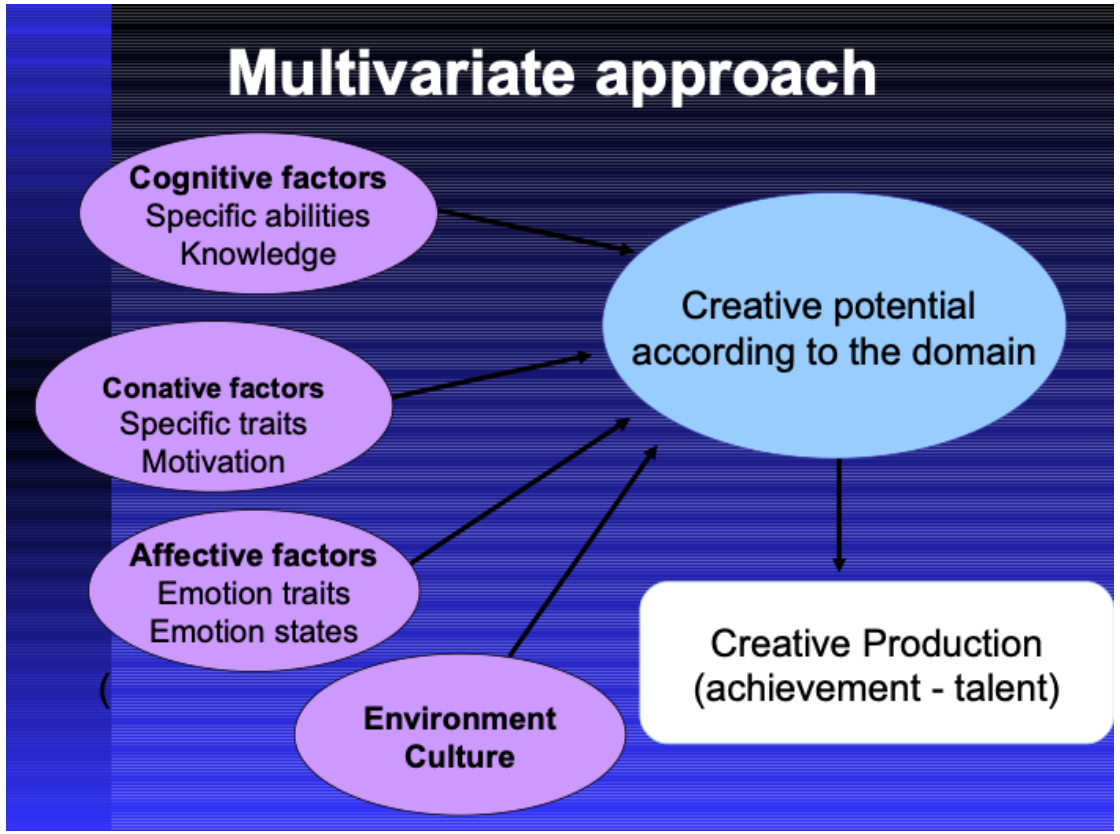
Expressional fluency (FE): The ability to rapidly think of different ways of expressing an idea.

Figural fluency (FF): The ability to rapidly draw or sketch as many things (or elaborations) as possible when presented with a nonmeaningful visual stimulus (e.g., a set of unique visual elements).

Figural flexibility (FX): The ability to rapidly draw different solutions to figural problems.

What is Creativity?

The Investment Theory of Creativity by Robert J. Sternberg and Todd I. Lubart (1991)



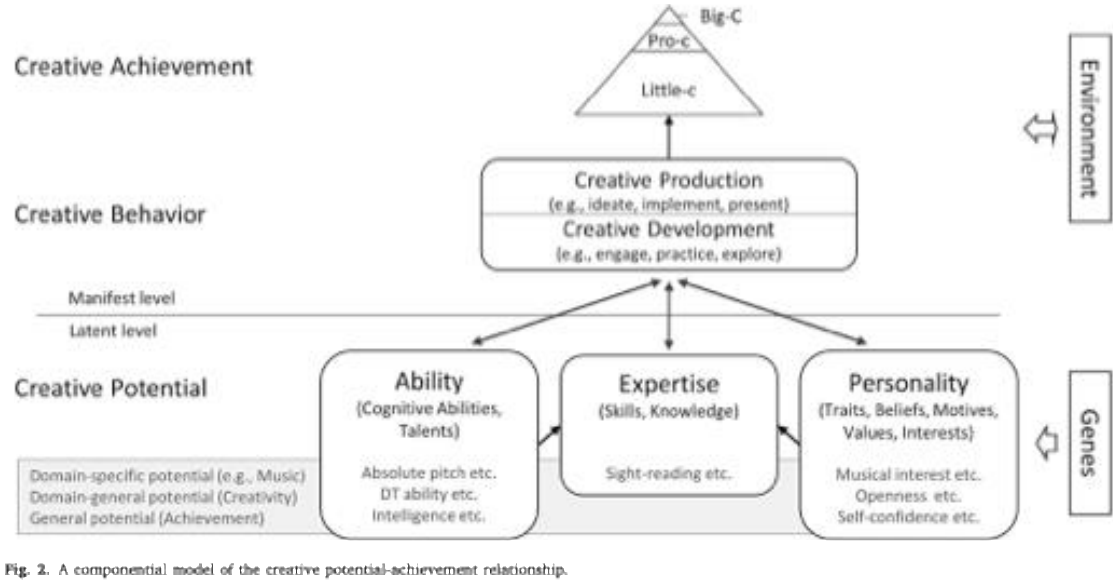
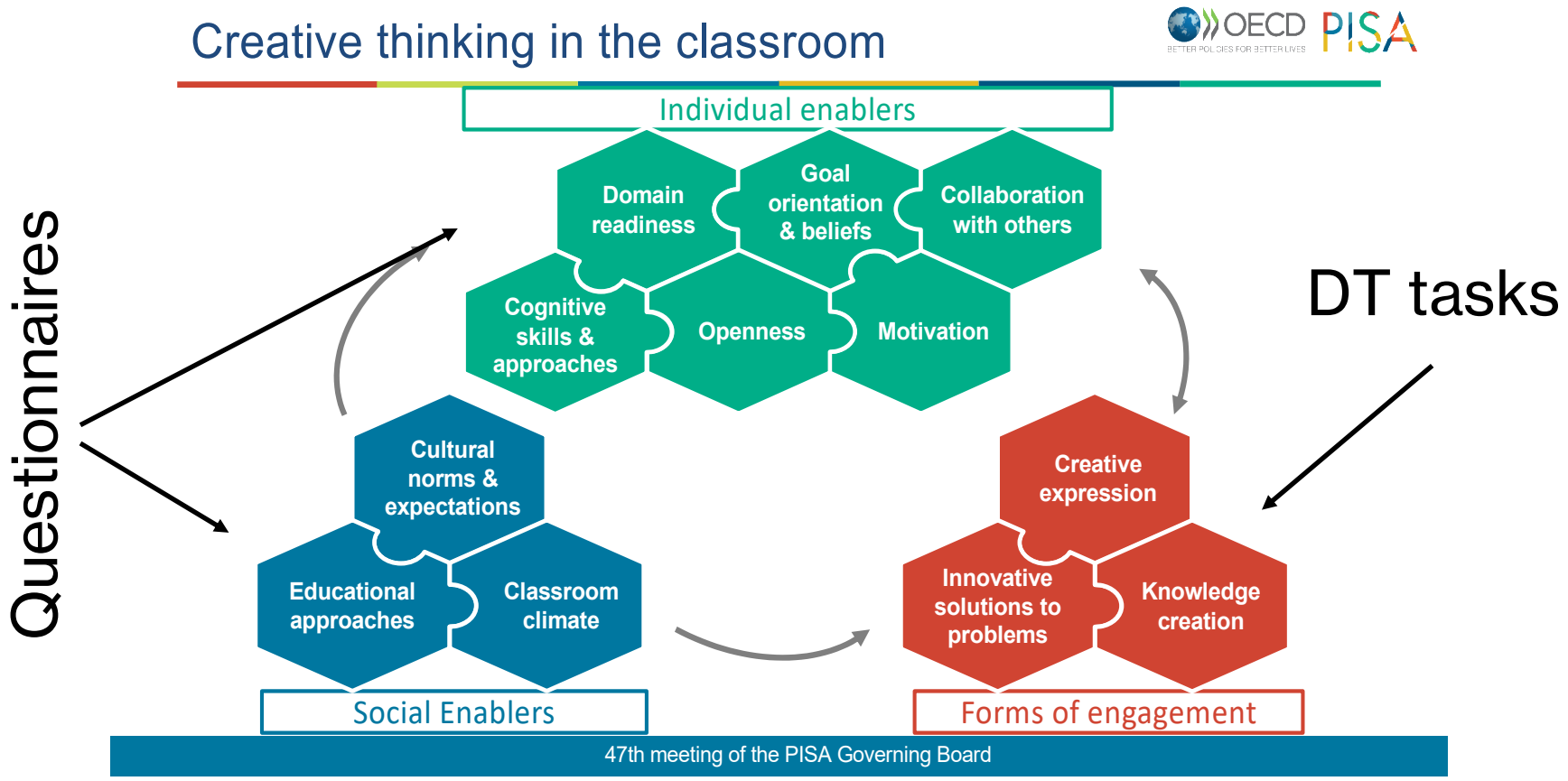


Fig. 2. A componential model of the creative potential-achievement relationship.

Domain specific

PISA Framework

Inovative domain: Creativity Thinking



Creativity Assessment

Creativity and Arts Tasks and Scales: Free for Public Use (Silvia & Benedek, 2023)

https://osf.io/4s9p6/

Divergent thinking tasks

Alternative Uses Task (AUT)

For this task, you'll be asked to come up with as many original and creative uses for a BRICK as you can. The goal is to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different.

Your ideas don't have to be practical or realistic; they can be silly or strange, even, so long as they are CREATIVE uses rather than ordinary uses.

You can enter as many ideas as you like; The task will take 3 minutes.

You can type in as many ideas as you like until then, but creative quality is more important than

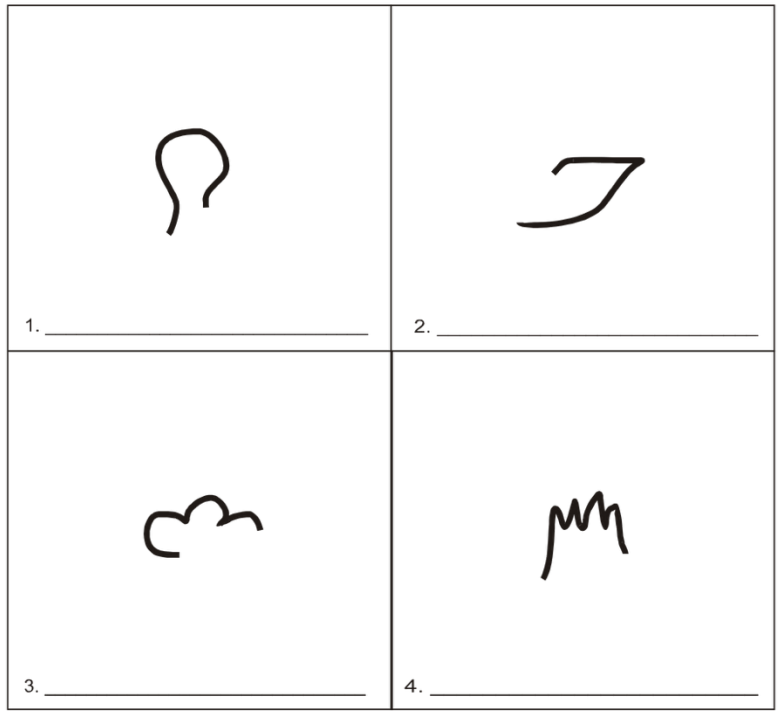
quantity. It's better to have a few really good ideas than a lot of uncreative ones.

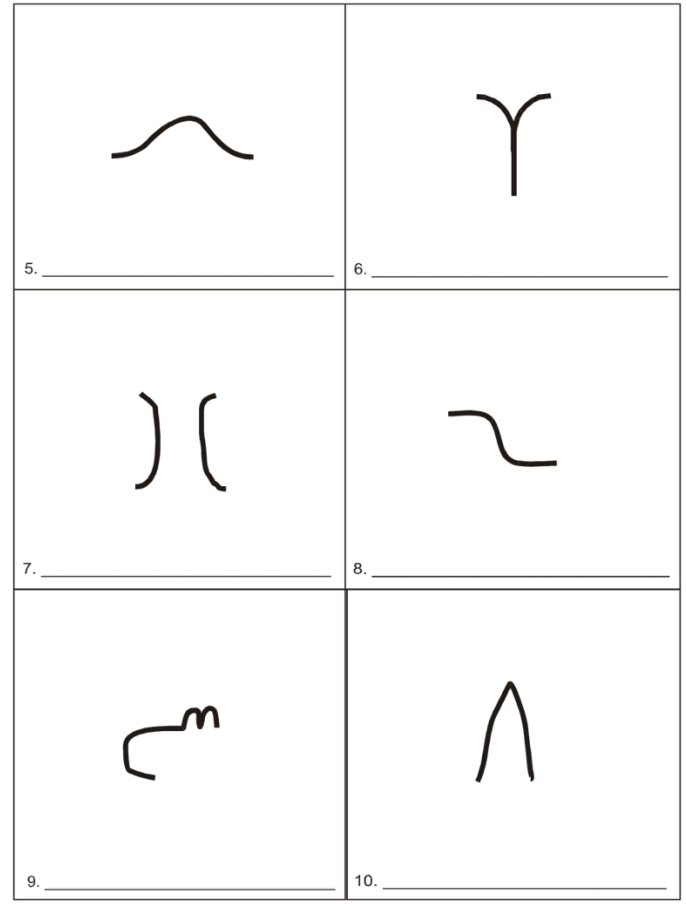
Please list all of the creative, unusual uses for a BRICK you can think of.

- _____
- _____
- _____
- _____

ATIVIDADE 2 – COMPLETANDO AS FIGURAS

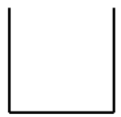
Se você olhar as 10 figuras a seguir perceberá que, se você completá-las, poderá fazer desenhos interessantes e bastante completos. Olhe para cada uma e imagine qual desenho pode surgir a partir delas. Lembre-se que as figuras devem fazer parte do seu desenho. Tente pensar em alguma idéia que você acha que as outras pessoas não pensariam. Quando terminar de desenhar invente um título para cada desenho e escreva-o na linha ao lado do número de cada figura.



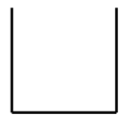


ATIVIDADE 3 – CRIANDO FIGURAS

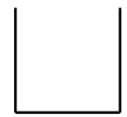
Nessa atividade sua tarefa é realizar desenhos usando as figuras das próximas três páginas. Pense em coisas que você conhece que possuem este formato. Você pode desenhar onde quiser, mas sempre lembrando-se que as figuras devem fazer parte do seu desenho. Novamente tente pensar em idéias diferentes, que você acha que ninguém mais pensaria. Tente fazer com que seus desenhos sejam bastante completos e interessantes. Quando terminar de desenhar, invente um título e escreva-o na linha numerada abaixo de cada figura.



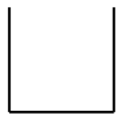
1. _____



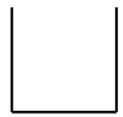
2. _____



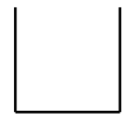
3. _____



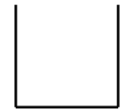
4. _____



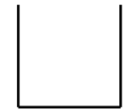
5. _____



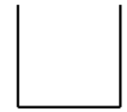
6. _____



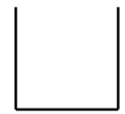
7. _____



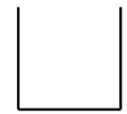
8. _____



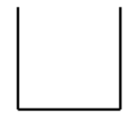
9. _____



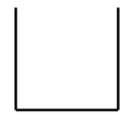
10. _____



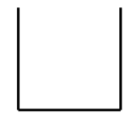
11. _____



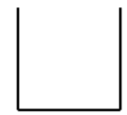
12. _____



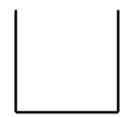
13. _____



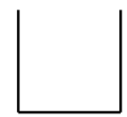
14. _____



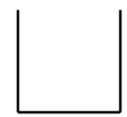
15. _____



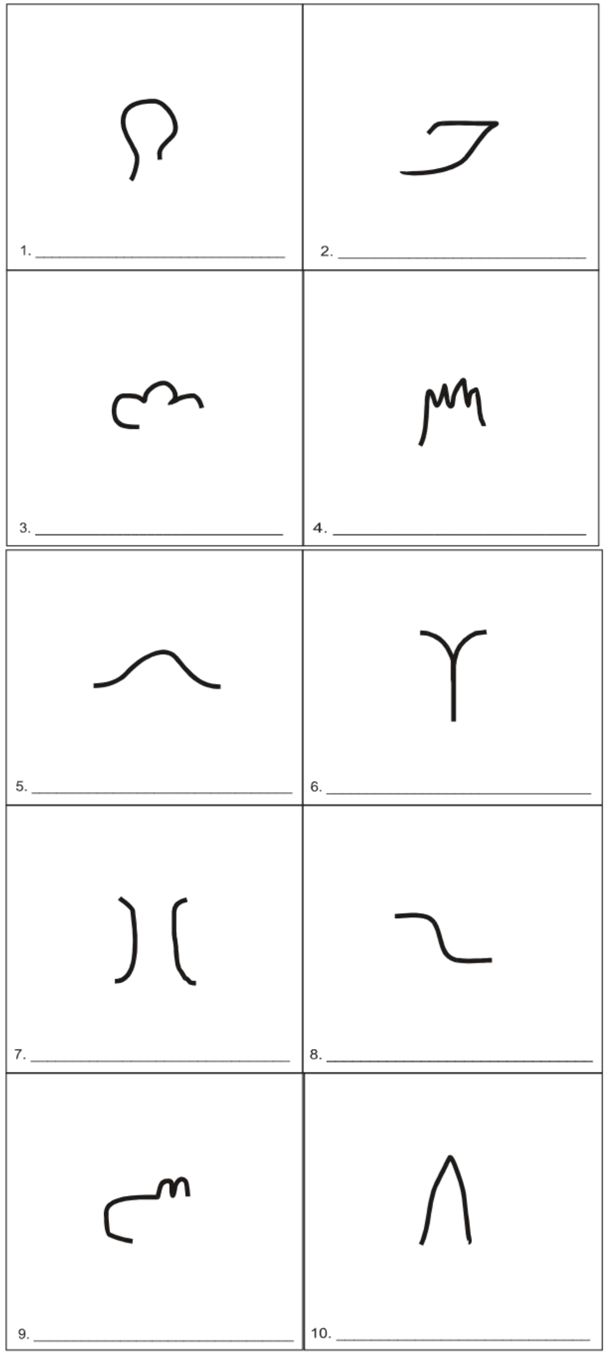
16. _____



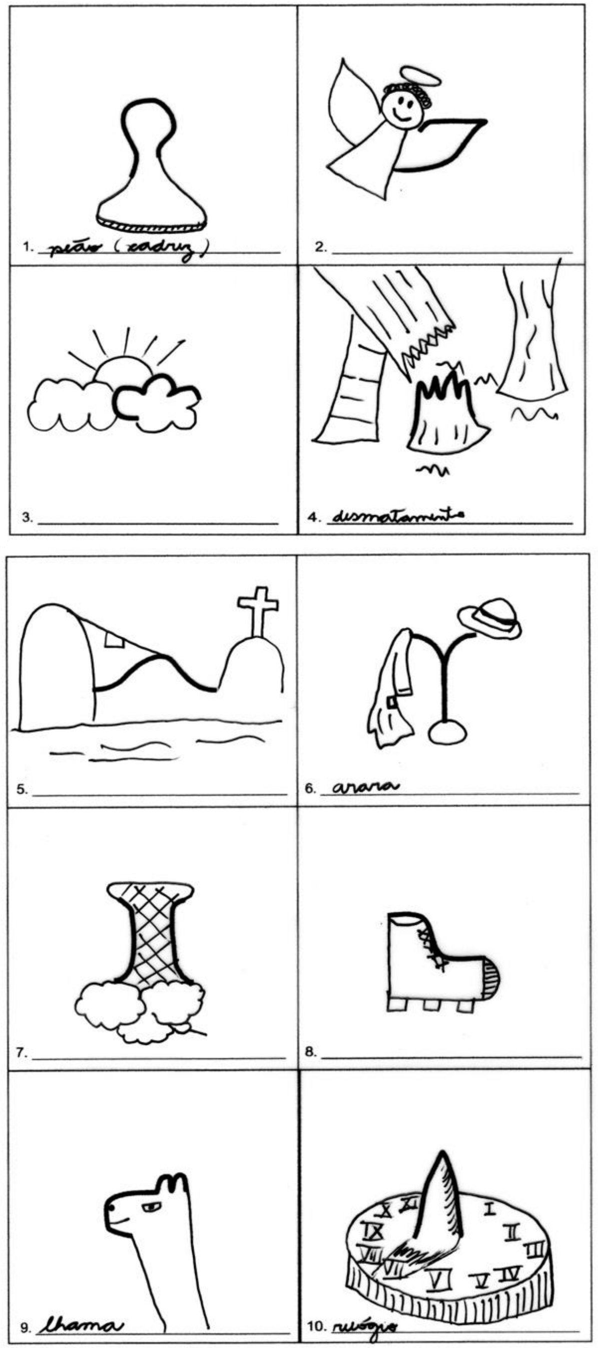
17. _____



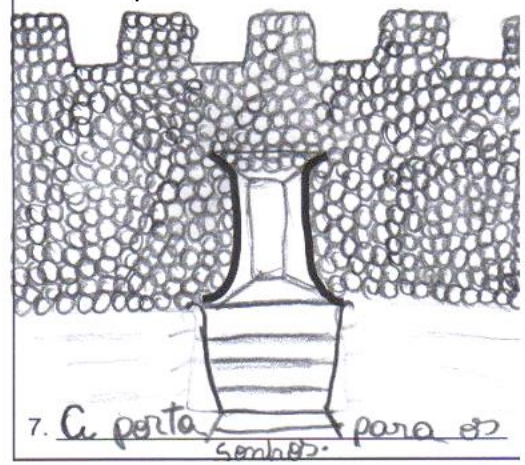
18. _____

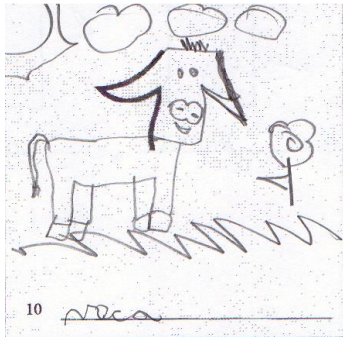


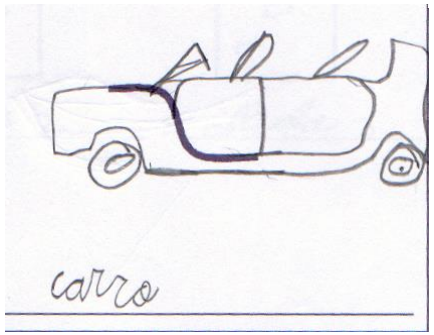
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.



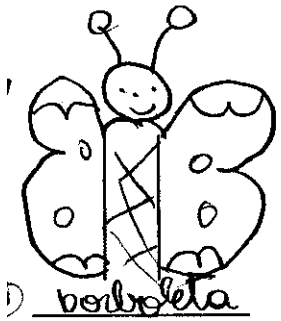
1. peão (xadrez)
2.
3.
4. desmatamento
5.
6. arara
7.
8.
9. lhama
10. relógio

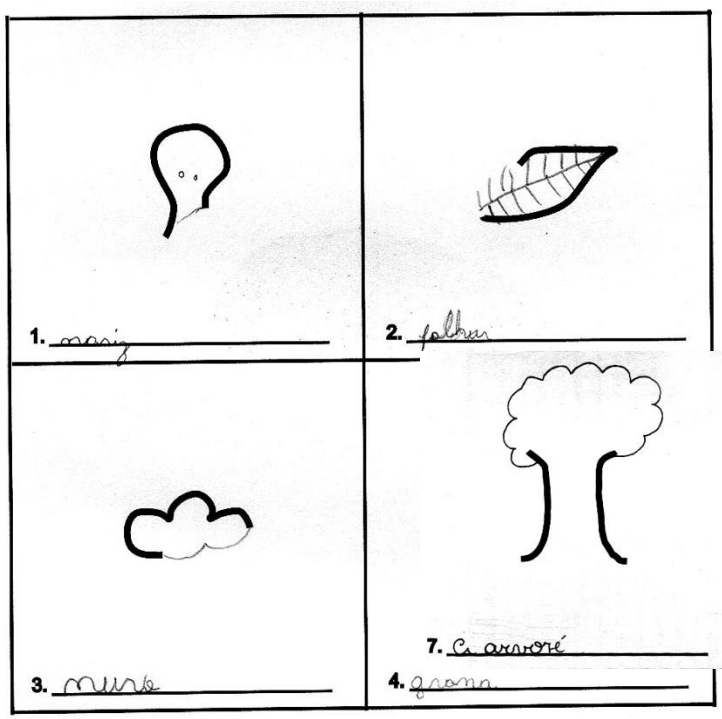
Baixa criatividade	Alta criatividade
 7. árvore.	 7. A porta para os sonhos.
“Árvore”	“A porta para os sonhos”
 7. Rua	 7. O amor está no ar
“Rua”	¨O amor está no ar¨

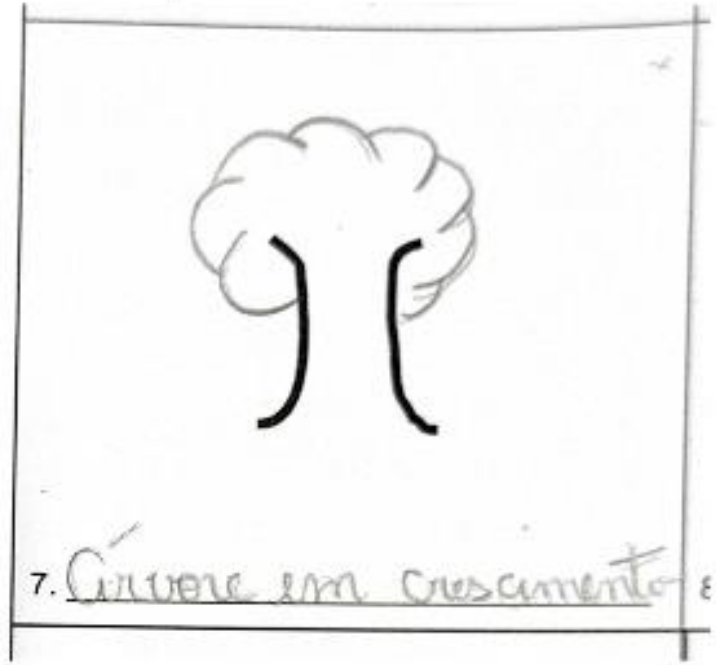

10
vaca

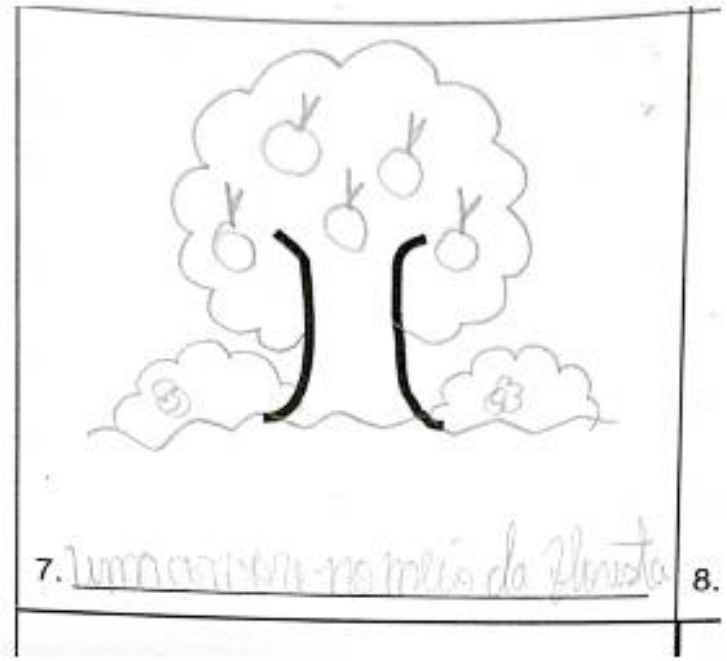

carro

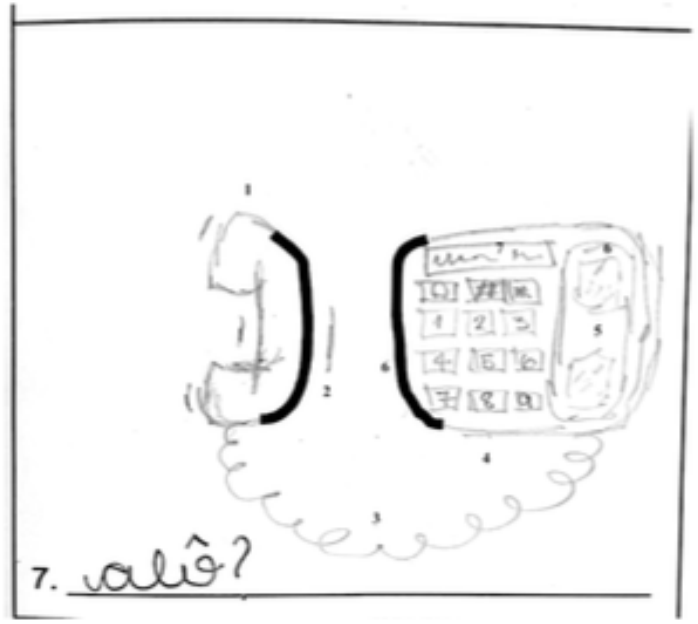

7. O jardim


borboleta

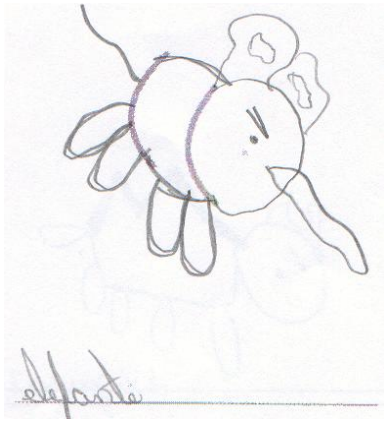

1. nariz
2. folha
3. nuvem
7. Ca arvoré
4. grama


7. Árvore em crescimento 8.

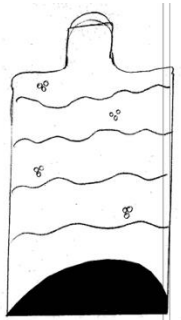

7. Uma macieira no início da Primavera 8.


7. olá?

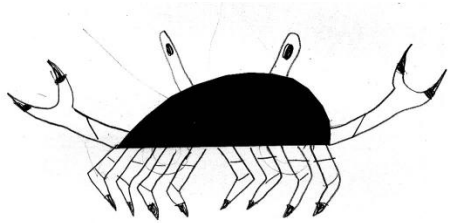
1. mala de corra



elefante

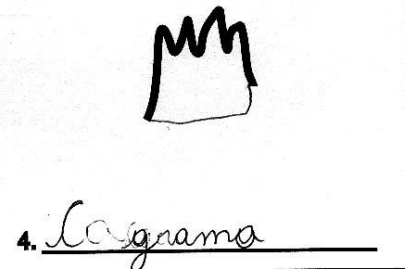


Título: girafa



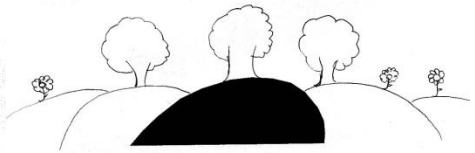
Título: O rinri no deserto

1. A calteça



4. O grama

3. As nuvem



Título: A paisagem

Criteria for Assessment (rubrics for raters)

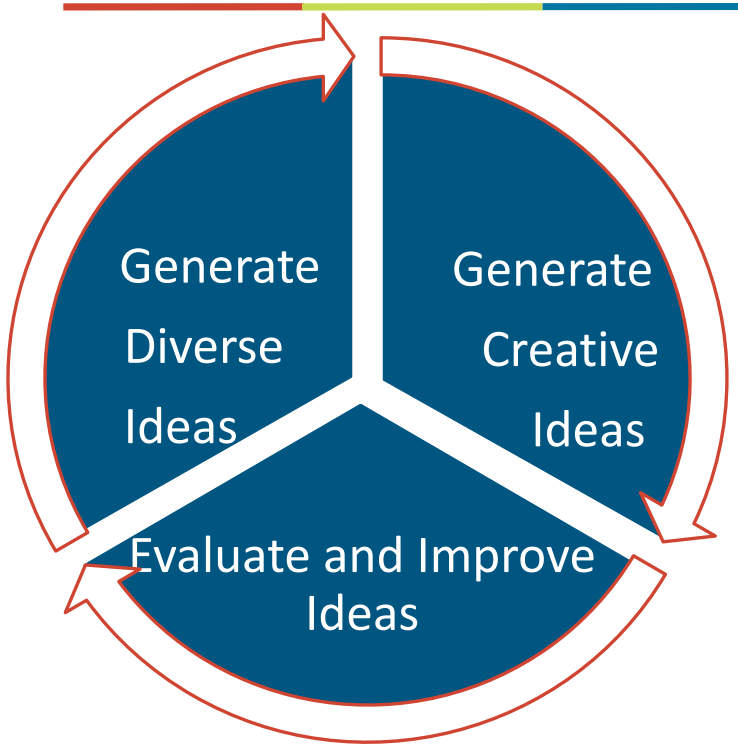
Reiter-Palmon, Forthmann, & Barbot (2019)

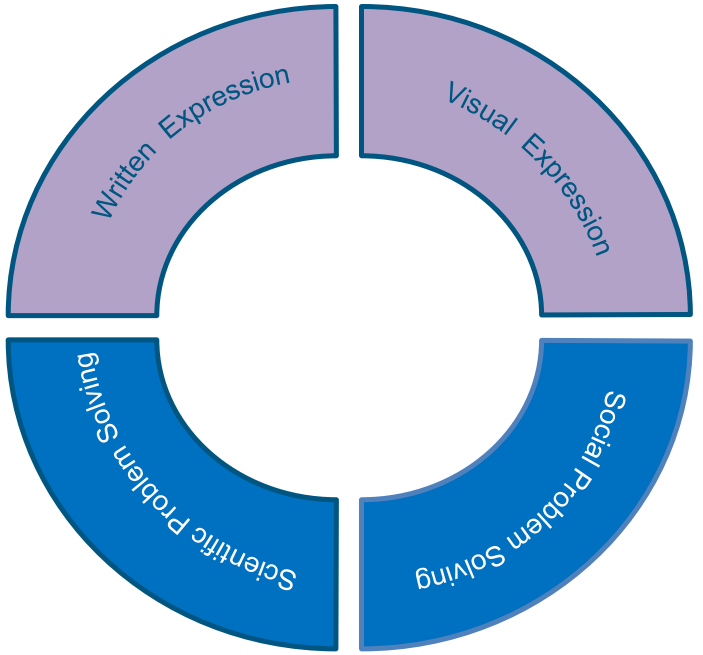
Criterion	Definition
Appropriateness	Fit, feasibility, and relevance to the task
Usefulness	Practical value and applicability
Fluency	Quantity of ideas produced
Originality	Novelty and uniqueness of ideas
Flexibility	Diversity and range of conceptual categories

- Wilson, Guilford and Chistesen (1953) defined originality in three ways: uncommonness, remote associations and cleverness as evaluated by ratings
- More objective ways (fluency and originality) vs More “subjective” ways Silvia et al. (2008)

- Holistic Judgment Method
- Criterion based Method
- Per response vs Per batch vs Top 2 or 5

Competency model for PISA CT





Written Expression

Task 1/3

DICE AND STORIES

Refer to the image on the right.

You are playing a game in which you have to roll dice and then connect the images that appear face up as inspiration for a story. As a warm-up you are using only two dice.

Create 2 different stories that connect the images to the right. The story ideas should be as different from each other as possible.

We recommend that you spend no longer than 7 minutes on this question.

Story 1

Story 2

Tested facet ‘generate diverse ideas’.
Possible scores: 0,1 2
Full credit if ideas are **appropriate (**if they make reference to both images in the dice) and **different** (distinct categories)

Written Expression

Task 2/3

DICE AND STORIES

Refer to the image on the right.

Now that you have some practice with the game, try to write one creative story that connects the six images on the right in the order they appear. Your story will receive a high score if it is original, demonstrates a rich imagination and is well structured.

We recommend that you spend no longer than 5 minutes on this question, and use no more than 80 words.

Write your story here

Tested facet ‘generate creative ideas’.
Possible scores: 0,1 2
Full credit if ideas are **appropriate (**if they make reference to all images in the dice) and **original** (unconventional theme or approach, not present in lists of conventional responses)

Written Expression

Task 3/3

Refer to the image and text on the right.

Now you are playing a variation of the game in which you create a story with a friend. Read the start of the story your friend has written using the six images in the top panel. You have to continue the story using the three images in the bottom panel.

Write a creative continuation of your friend's story trying to follow his inspiration and style.

We recommend that you spend no longer than 5 minutes on this question, and use no more than 80 words.

DICE AND STORIES

The six images used by your friend

Your friend's story

With a heavy heart, it was time to say goodbye. Her house no longer felt like home. She packed up her things and headed East; across the globe to China. She'd never been afraid of new challenges and she was ready to roll the dice and see where it took her next.

The three images you have to use to continue the story

Write here

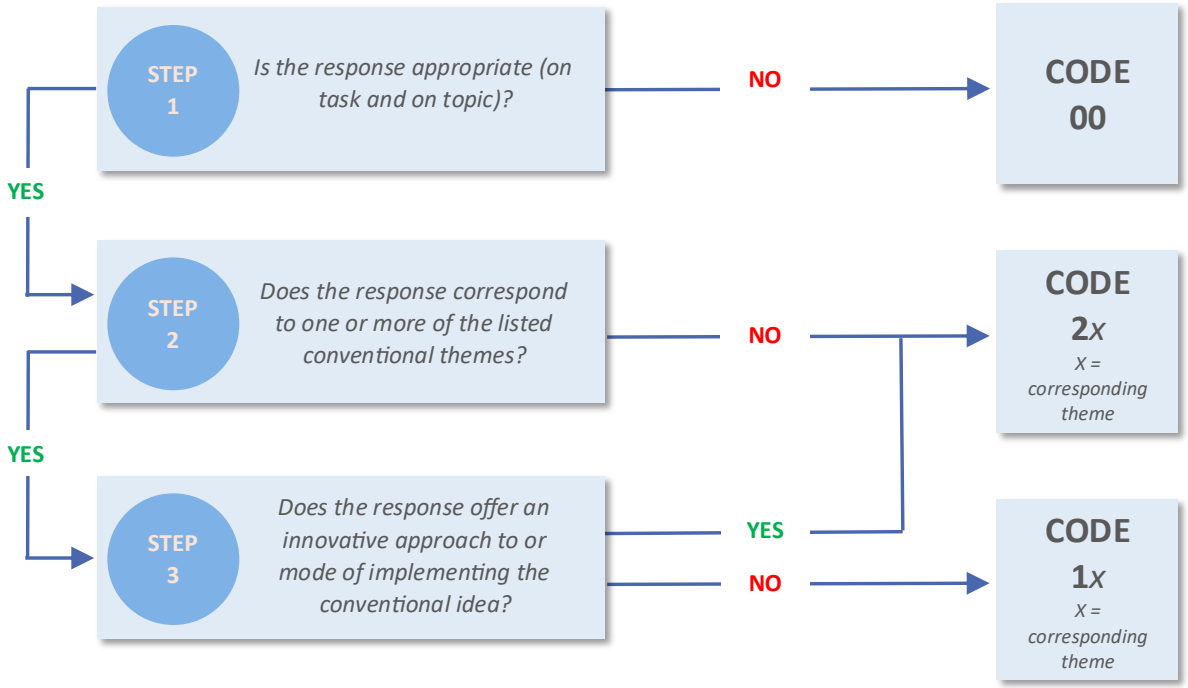
Tested facet 'Evaluate and improve ideas'.
Possible scores: 0,1 2
Full credit if ideas are **appropriate (**coherent iteration of an existing idea) and **original** (not present in lists of conventional responses)

Sample item

Written expression / Generate Creative Ideas

General Coding Guides

Coding process for **Generate Creative Ideas** (GCI) AND **Evaluate & Improve Ideas** (EII)



Creative Achievement

Literature

Music

Arts and crafts

Creative Cooking

Sports

Visual Arts (Graphics, Painting, Sculpting, Architecture)

Performing Arts (Theatre, Dance, Film)

Science & Engineering

Top 5 Creative Achievments

Inventory of Creative Activities and Achievements (ICAA)

The ICAA was devised by Benedek, M. & Jauk, E. (2012) at the University of Graz, Austria.

It comprises a creative achievement scale (8 domains) and a creative activities scale (same 8 domains), which are presented together for each domain. Additionally, it includes an optional top-5 creative achievement list.

The achievement scale is scored similarly to the CAQ by Carson.

The ICAA was used and described in:

- Jauk, E., Benedek, M., & Neubauer, A. C. (in press). The road to creative achievement: A latent variable model of ability and personality predictors. *European Journal of Personality*. doi:10.1002/per.1941
- Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence, 41*, 212-221. doi:10.1016/j.intell.2013.03.003

Creative Achievement

I. Literature

*1) For each activity in the domain of **literature**, please indicate how often you have carried out this activity over the past 10 years.* Mark the applicable box.

	Never	1-2 times	3-5 times	6-10 times	More than 10 times
Wrote a short literary work (e.g., poem, short story)					
Wrote a long literary work (e.g., book, theatre play)					
Wrote a newspaper article/editorials					
Created an original talk					
Made up a joke					
Wrote a blog entry					

*2) Please mark all levels of achievement that you have attained in the domain of **literature**:*

- o 0. I have never been engaged in this domain
- o 1. I have tried this domain once.
- o 2. I have already created at least one original work in this domain.
- o 3. I have presented my original work in this domain to some friends.
- o 4. I have presented my original work in this domain to strangers.
- o 5. I have already taken classes to improve my skills in this domain.
- o 6. I have already published my original work in this domain.
- o 7. I have already participated in a contest in this domain.
- o 8. I have already won an award or prize for my original work in this domain.
- o 9. Media have already reported about my work in this domain.
- o 10. I have already sold some of my work in this domain.

*3) Please state for how many years of your life (approximately) you have been engaged in the domain of **literature**:*

_____ years

II. Music

*1) For each activity in the domain of **music**, please indicate how often you have carried out this activity over the past 10 years. Mark the applicable box.*

	Never	1-2 times	3-5 times	6-10 times	More than 10 times
Wrote a piece of music					
Reinterpreted a piece of music in a creative way					
Made up a melody					
Made up a rhythm					
Artificially created sounds (e.g., via computer or synthesizer)					
Created a mix-tape (or any other compilation of songs; e.g., DJ-ing)					

*2) Please mark all levels of achievement that you have attained in the domain of **music**:*

- o 0. I have never been engaged in this domain
- o 1. I have tried this domain once.
- o 2. I have already created at least one original work in this domain.
- o 3. I have presented my original work in this domain to some friends.
- o 4. I have presented my original work in this domain to strangers.
- o 5. I have already taken classes to improve my skills in this domain.
- o 6. I have already published my original work in this domain.
- o 7. I have already participated in a contest in this domain.
- o 8. I have already won an award or prize for my original work in this domain.
- o 9. Media have already reported about my work in this domain.
- o 10. I have already sold some of my work in this domain.

*3) Please state for how many years of your life (approximately) you have been engaged in the domain of **music**:*

_____ years

Challenges

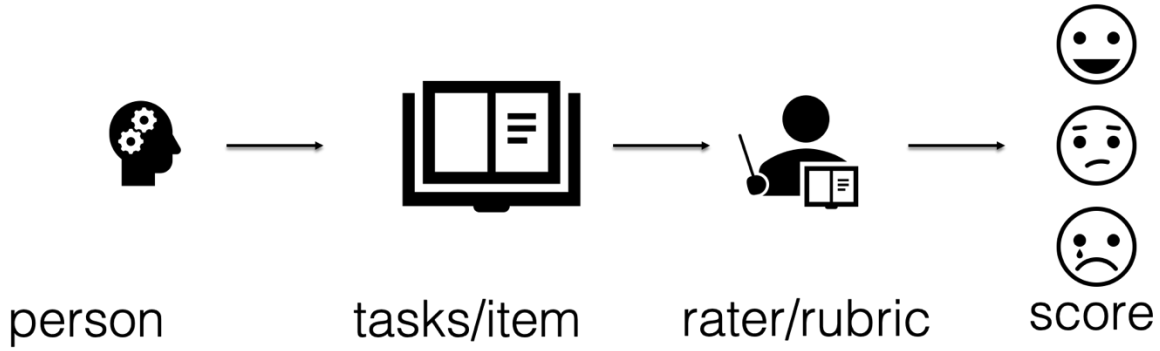
- Uniqueness (statistical infrequency) is sample dependent
  - In small samples a responses is more likely to be scored for uniqueness (more creative) than if it were scored in large samples

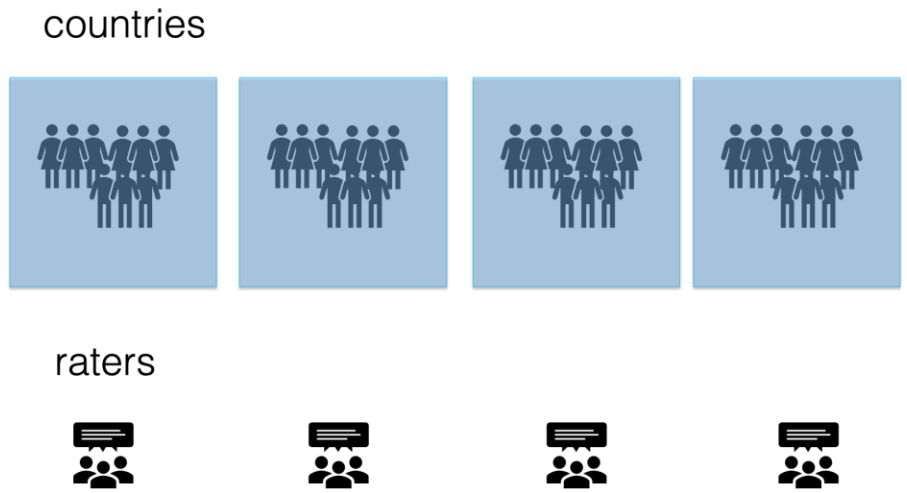
- Uniqueness, fluency and flexibility are highly correlated (confounded). $r > .88$
  - Fluency = number of categories (flexibility) + number of repetitions within categories
  - Flexibility = Fluency - number of repetitions within categories (method covariance)

- Fluency is highly correlated with originality (confounded).
  - No independent measures of quantity and quality

Challenges

- Rater mediated assessment
- Costs of training and scoring sessions / workload on raters
- Cross cultural comparability (linking/equating tasks/ raters)
  - Multilevel structure





Psychometrics of divergent thinking tasks

Psychology of Aesthetics, Creativity, and the Arts

1931-3896/14/$12.00 http://dx.doi.org/10.1037/a0038055

Divergent Productions of Metaphors: Combining Many-Facet Rasch Measurement and Cognitive Psychology in the Assessment of Creativity

Ricardo Primi
University of São Francisco

This article presents a new method for the assessment of creativity in tasks such as "The camel is _____ of the desert." More specifically, the study uses Tourangeau and Sternberg's (1981) domain interaction model to produce an objective system for scoring metaphors produced by raters and the many-facet Rasch measurement to model the rating scale structure of the scoring points, item difficulty, and rater severity analysis, thus making it possible to have equated latent scores for subjects, regardless of rater severity. This study also investigates 4 aspects of the method: reliability, correlation between quality and quantity, criterion validity, and correlation with fluid intelligence. The database analyzed in this study consists of 12,418 responses to 9 items that were given by 975 persons. Two to 10 raters scored the quality and flexibility of each metaphor on a 4-point scale. Raters were counterbalanced in a judge-linking network to permit the equating of different "test forms" implied in combinations of raters. The reliability of subjects' latent quality scores was .88, and the correlation between quality and quantity was low ($r = -.14$), thus showing the desired separation between the 2 parameters established for the task scores. The latent score on the test was significantly associated with the profession that requires idea production ($r = .19$), and the latent scores for the correlation between creativity and fluid intelligence were high, $\beta = .51$, even after controlling for crystalized intelligence ($r = .47$). Mechanisms of fluid intelligence, executive function, and creativity are discussed.

Keywords: metaphor production, intelligence, creativity, item response theory, Rasch measurement

Instructions

In this test we want you to invent metaphors to complete sentences. See the examplesbelow:

The **camel** is a _____ of the **desert**

Metaphors	*Explanation*
1) *boat*	*In the sea the boat is a means of transport walking swinging like a camel in the desert*
2) *motorcycle*	*Because motorcycle is a transport for one or two people and need only a few amount of fuel like a camel in the desert that needs little water*
3) *slug*	*Because his walking is slowly, marking the floor and swung her butt like the camel*
4) *Barrichello*	*Because when he is not stopped he is walking slowly*

Nine items (1 to 4 responses asked):
The horn is the _____ of the car
The grass is the _____ of the land
The stars are the_____ of the night
The Ball is the _____ of the players
The planets are the _____ of the sun
The bus are the _____ of the city
The hanger is the _____ of the clothes
The door is the _____ of the house
The fish are the _____ of the sea

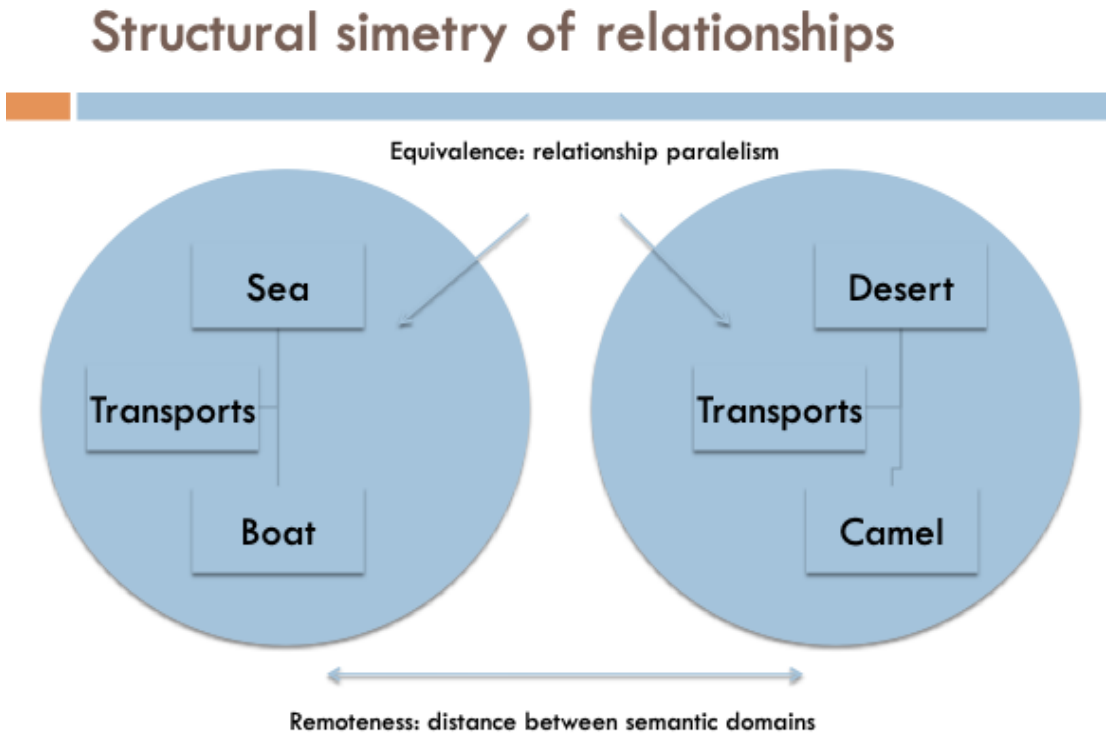
Cognitive Psychology of Metaphors

- Metaphors, are means of expressing different characteristics of a concept (camel, hanger, cat's mustache) through association of ideas.
  - "The camel is the boat of the desert",
  - The hanger is the clothes' spinal cord",
  - The mustache is the antenna of the cat"
- Underlying metaphors creation there are basic cognitive processes, like analogical reasoning and remote associations that are linked to creativity
- The fact that metaphors is an ideational product that has a remote association embedded in it make it useful resource to be used in creativity assessment

Tourangeu and Sternberg's (1981) domain interaction theory

- Metaphors reclassify the meaning of one object/event/idea seeing it thought the lens of another domain (and its relationships) based on structural similarity
  - Tenor: subject or topic of the metaphor (camel)
  - Vehicle: predicate of the sentence that contains new information about the tenor (boat)
  - Ground: relational base / domain / memory structure
- The act of seeing something using another domain change the meaning of the features within each domain, therefore, it is called a domain interaction.
- Tourangeau & Sternberg (1981) found that the the aptness of metaphor is related
  - to the degree to which concepts occupy similar positions with respect to their own domain
  - to the distance between domains (inverted u-shaped)
- In this theory good metaphors have equivalent or parallel relationships across domains (equivalence) and the domains are semantically distant (remoteness)
- Therefore good metaphors are clear and have a optimal level of remoteness

Structural elements of metaphors



Analogies	Metaphors
A is to B as C is to D A:B // C:D	A is the C of B A (// C:?) : B
Camel: Desert / / Boat: Sea	The camel is the ship of the desert

COGNITIVE PSYCHOLOGY 5, 1–28 (1973)

A Model for Analogical Reasoning[1]

DAVID E. RUMELHART AND ADELE A. ABRAHAMSON
University of California, San Diego

A theory of analogical reasoning is proposed in which the elements of a set of concepts, e.g., animals, are represented as points in a multidimensional Euclidean space. Four elements A,B,C,D, are in an analogical relationship A:B::C:D if the vector distance from A to B is the same as that from C to D. Given three elements A,B,C, an ideal solution point I for A:B::C:? exists. In a problem A:B::C:D_1, . . ., D_i, . . ., D_n, the probability of choosing D_i as the best solution is a monotonic decreasing function of the absolute distance of D_i from I. A stronger decision rule incorporating a negative exponential function in Luce's choice rule is also proposed. Both the strong and weak versions of the theory were supported in two experiments where Ss rank-ordered the alternatives in problems A:B::C:D_1,D_2, $D_3$$D_4$. In a third experiment the theory was applied and further tested in teaching new concepts by analogy.

To introduce our notion of analogical reasoning, it is useful to outline a definition of the word *reasoning* from which we can work. The term is used here to denote those processes in information retrieval which depend on the *structure*, as opposed to the *content* of organized memory.

Thus, one might answer the question "Who is the father of your country?" in at least two different ways. In one case, the specific information that George Washington was the "father of our country" might be stored and used to answer the question. On the other hand, when specific information is not available, one can consult the stored meanings of the words in question and one's knowledge of history to derive a plausible answer. The first of these methods might be called remembering, since retrieval depends on the specific information stored. The second method may be identified with reasoning, since in this case retrieval depends to a much greater extent on the *form* of the relationship among the words. The same act of reasoning (i.e., the same processes) could have been applied to the question "Who was the father of your state?" or "Who was the mother of your country?". It is not the specific content of the question but the form of the relationships among the words which determines the response.

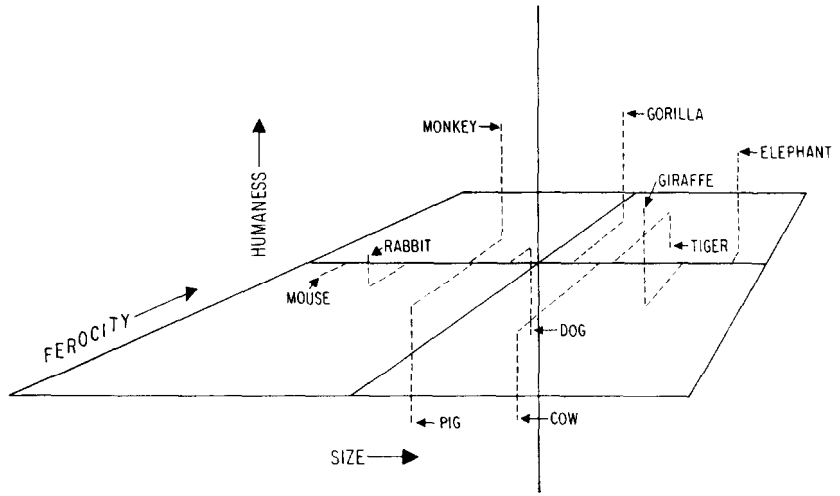


FIG. 1. The placements of a selected set of animals based on data from Henley (1969).

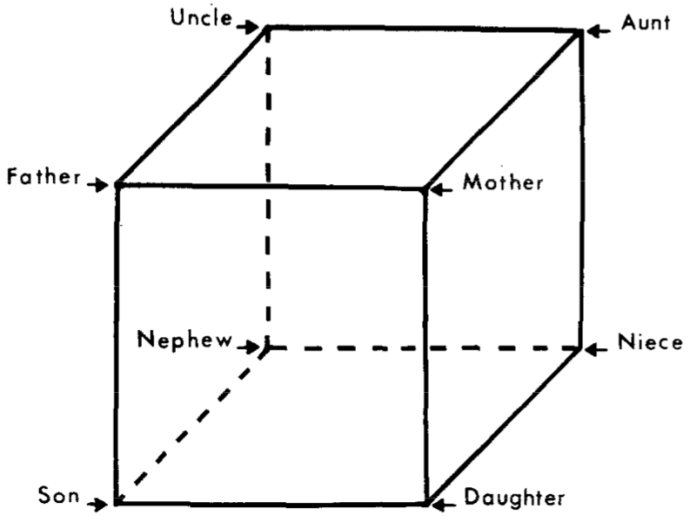


FIG. 2. Three-dimensional representation of relations among eight kinship terms.

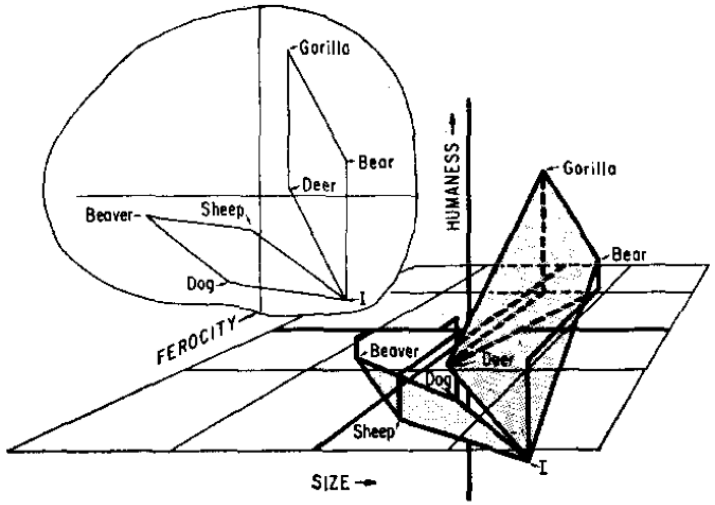


FIG. 5. Three-dimensional representation of two analogies with the same solution. The analogies are (a) GORILLA:DEER::BEAR:? and (b) BEAVER:SHEEP::DOG:?.

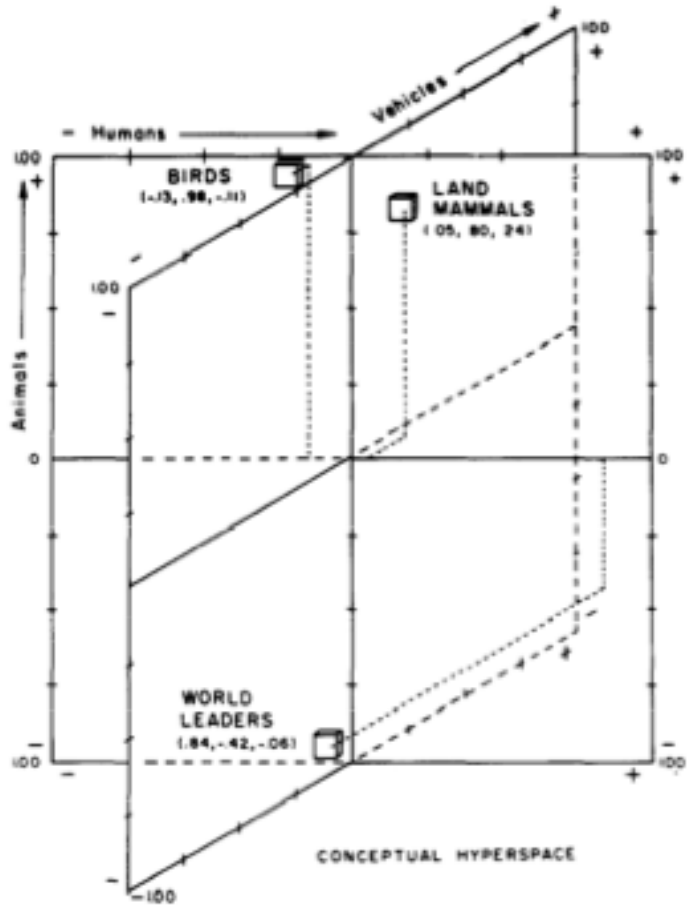
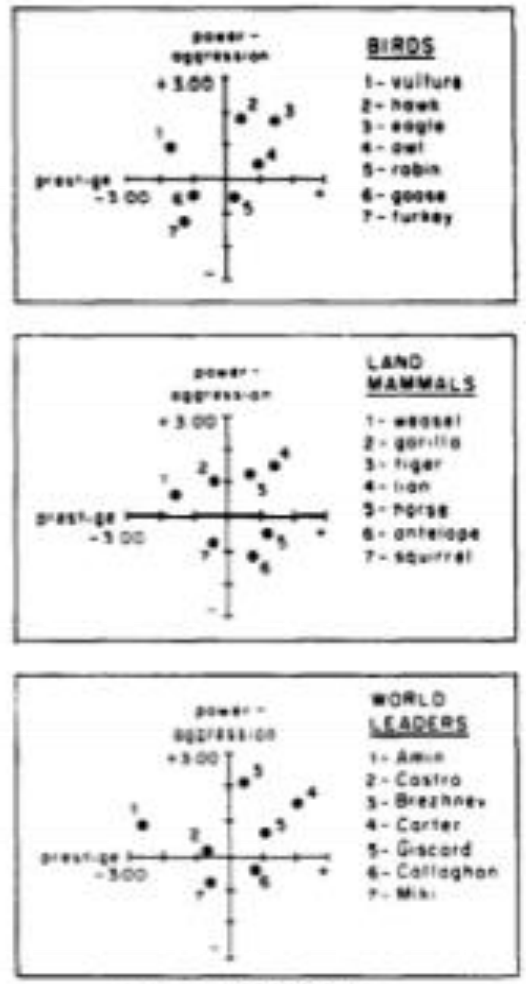


FIG. 1. Relation of higher-order and lower-order spaces. To the left is shown the space of domains, or hyperspace. Each point in this space is itself a full space of a lower order, as shown to the right. For example, the point "birds" in the hyperspace maps into the space of birds at the top right of the figure.



LOWER ORDER SPACE

FIGURE 1—*Continued*

Metaphor Creation Test (MCT): the task

Instructions

In this test we want you to invent metaphors to complete sentences. See the examplesbelow:

The **camel** is a _____ of the **desert**

Metaphors	*Explanation*
1) *boat*	*In the sea the boat is a means of transport walking swinging like a camel in the desert*
2) *motorcycle*	*Because motorcycle is a transport for one or two people and need only a few amount of fuel like a camel in the desert that needs little water*
3) *slug*	*Because his walking is slowly, marking the floor and swung her butt like the camel*
4) *Barrichello*	*Because when he is not stopped he is walking slowly*

- Nine items (1 to four responses asked):
- The horn is the _____ of the car
- The grass is the _____ of the land
- The stars are the_____ of the night
- The Ball is the _____ of the players
- The planets are the _____ of the sun
- The bus are the _____ of the city
- The hanger is the _____ of the clothes
- The door is the _____ of the house
- The fish are the _____ of the sea

Scoring Criteria

- Judges rates each idea in a 4-point scale:
  - **0**: not a metaphor
  - **1**: correct metaphor (equivalent and remote in a average degree)
  - **2**: correct metaphor (equivalent and remote in a advanced degree, explanation make the underlying analogy clear, mapping the concepts)
  - **3**: correct metaphor (same as 2 but remote in a very advanced degree, we may consider the “wow! effect”)

Psychometric Modeling

Many Faceted Rasch Model (MFFM, Linacre, 1989)

In a three-facet MFRM model (subjects, items and raters),

the odds of a subject *n* receiving a rating *k* versus an immediate lower rating *k* -1 on item *i* by rater *j* is

$$log \frac{P_{nijk}}{P_{nijk-1}} = B_n - D_i - C_j - F_k$$

P_{nijk} and P_{nijk-1} refer to probabilities of receiving *k* or *k*-1 points;

B_n refers to subject n ability;

D_i refers to item *i* difficulty;

C_j refers to rater *j* severity;

F_k refers to rubric *k* level difficulty.

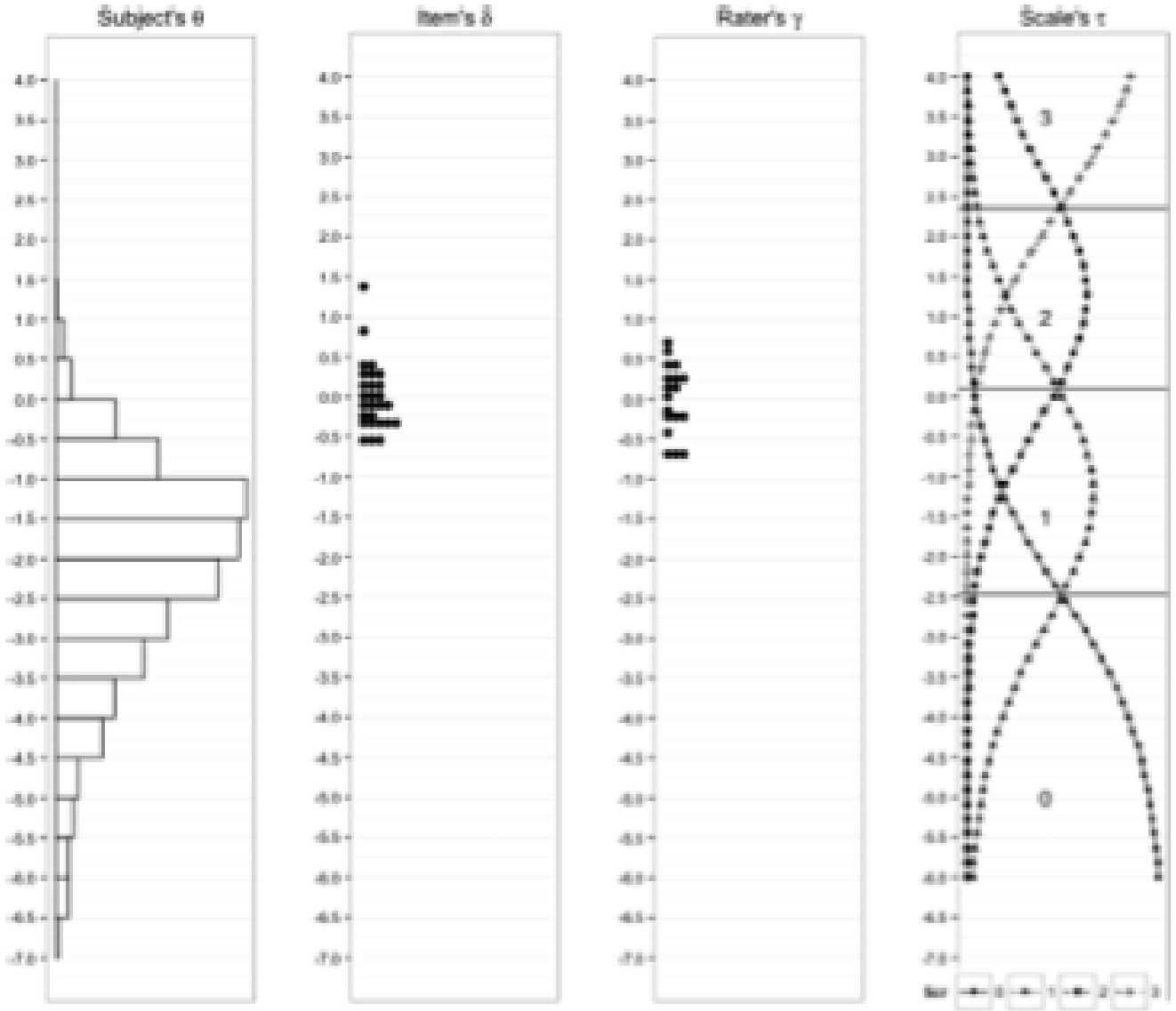
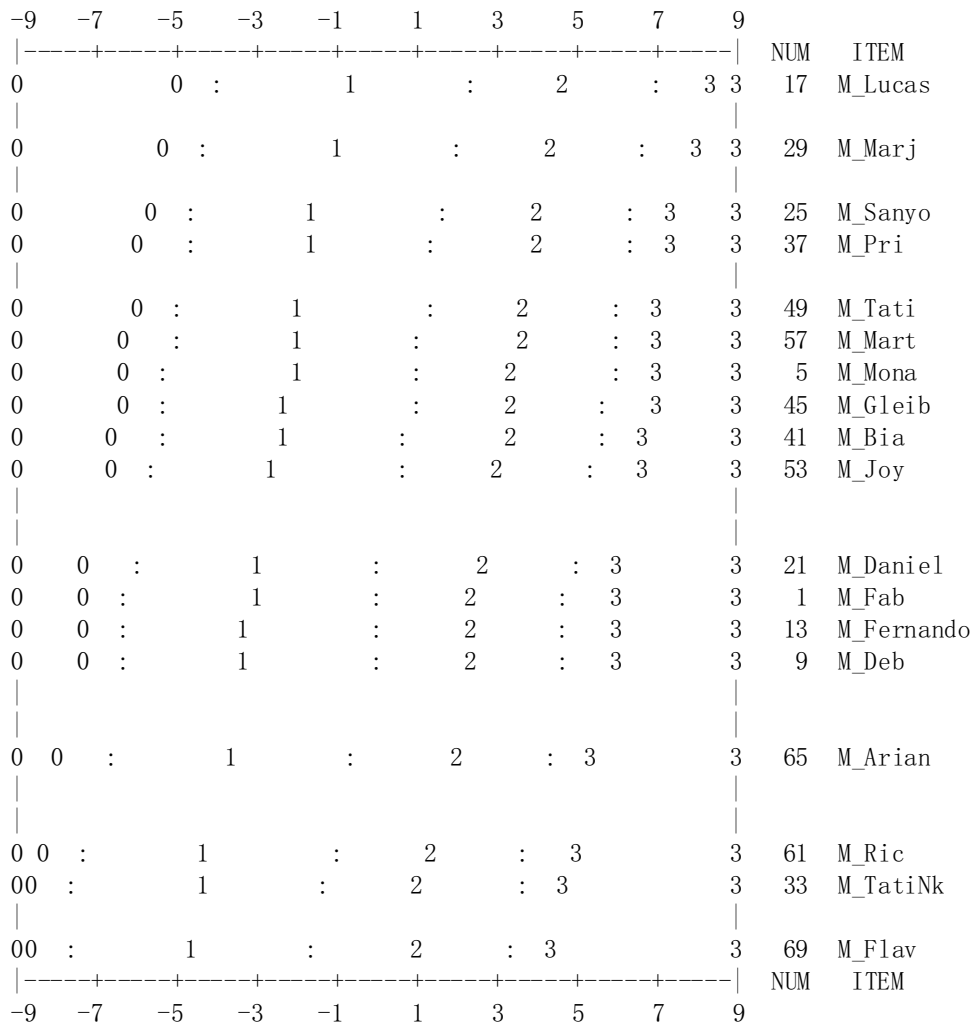


Figure 2. Construct maps showing the relationships of the main facets: persons, items, raters, and scale structure. θ = subject's theta; δ = item's delta; γ = raters' gamma; τ = scale's threshold.



1

221797322325212 14608431 23242 121

257535831364432583276906981092084480421 2

445064262097391564849607643081411808049 0 66201 8622 PERSON

S M S T

0 10 20 30 40 50 70 80 90 99 PERCENTILE

Table 1

Raters Model Parameters From MFRM and Their Model Fit

Rater (j)	Average score	Fair average score	γ_j	*SE* of γ_j	Infit	Outfit	r_{rt} (Meas)	r_{rt} (Exp)
R01	0.51	0.46	0.71	0.07	0.77	0.74	0.52	0.47
R02	0.47	0.49	0.60	0.04	1.06	1.08	0.44	0.49
R03	0.63	0.55	0.43	0.04	0.87	0.95	0.46	0.49
R04	0.59	0.55	0.42	0.05	0.93	0.88	0.59	0.53
R05	0.81	0.59	0.29	0.04	0.86	0.89	0.50	0.51
R06	0.86	0.59	0.29	0.06	0.73	0.75	0.32	0.46
R07	0.69	0.61	0.23	0.04	0.71	0.75	0.51	0.47
R08	0.83	0.63	0.16	0.05	0.84	0.91	0.46	0.50
R09	0.88	0.64	0.13	0.04	1.16	1.15	0.57	0.56
R10	1.04	0.68	0.03	0.10	1.11	1.10	0.52	0.52
R11	0.87	0.73	−0.14	0.06	1.40	1.41	0.50	0.61
R12	0.77	0.75	−0.18	0.02	1.11	1.10	0.58	0.58
R13	0.85	0.76	−0.20	0.03	1.08	1.06	0.58	0.57
R14	0.82	0.77	−0.25	0.04	1.02	1.02	0.56	0.56
R15	0.98	0.84	−0.42	0.06	0.70	0.70	0.51	0.49
R16	0.69	0.93	−0.67	0.05	0.97	0.99	0.55	0.55
R17	1.32	0.94	−0.71	0.15	0.58	0.59	0.74	0.60
R18	1.04	0.94	−0.71	0.14	1.32	1.30	0.64	0.66

Note. γ_j gamma parameters indexing ater j severity, rrt rater j total correlations (Meas indicates observed and Exp indicates expected by the model).

Efficient Rating Designs

- Does every response need to be scored by every rater ? No.
- We just need a judge-linked design (also called a *connected* or *linked* design
- What is a linked design?
  - All raters (judges), examinees (students), and items (essays or prompts) are connected within a single measurement framework.
  - This connectivity is necessary to estimate the on a common scale.
  - Without this linkage, model parameters of various facets (judges, items, examinees) are isolated groups and therfefore would be measured on separate, incomparable scales

American Psychological Association
DIVISION 10
Psychology of Aesthetics, Creativity, and the Arts

1931-3896/19/$12.00

2019, Vol. 13, No. 2, 176–186
http://dx.doi.org/10.1037/aca0000230

Applying Many-Facet Rasch Modeling in the Assessment of Creativity

Ricardo Primi
Universidade São Francisco, Campinas, and Ayrton Senna Institute, São Paulo, Brazil

Paul J. Silvia
University of North Carolina at Greensboro

Emanuel Jauk
Technische Universität Dresden

Mathias Benedek
University of Graz

Creativity assessment with open-ended production tasks relies heavily on scoring the quality of a subject's ideas. This creates a faceted measurement structure involving persons, tasks (and ideas within tasks), and raters. Most studies, however, do not model possible systematic differences among raters. The present study examines individual rater differences in the context of a planned-missing design and its association with reliability and validity of creativity assessments. It applies the many-facet Rasch model (MFRM) to model and correct for these differences. We reanalyzed data from 2 studies (*N*s = 132 and 298) where subjects produced metaphors, alternate uses for common objects, and creative instances. Each idea was scored by several raters. We simulated several conditions of reduced load on raters where they scored subsets of responses. We then compared the reliability and validity of IRT estimated scores (original vs. IRT adjusted scores) on various conditions of missing data. Results show that (a) raters vary substantially on the lenient-severity dimension, so rater differences should be modeled; (b) when different combinations of raters assess different subsets of ideas, systematic rater differences confound subjects' scores, increasing measurement error and lowering criterion validity with external variables; and (c) MFRM adjustments effectively correct for rater effects, thus increasing correlations of scores obtained from partial with scores obtained with full data. We conclude that MFRM is a powerful means to model rater differences and reduce rater load in creativity research.

Keywords: creativity, assessment, many-facet Rasch model, raters, planned missing data

Supplemental materials: http://dx.doi.org/10.1037/aca0000230.supp

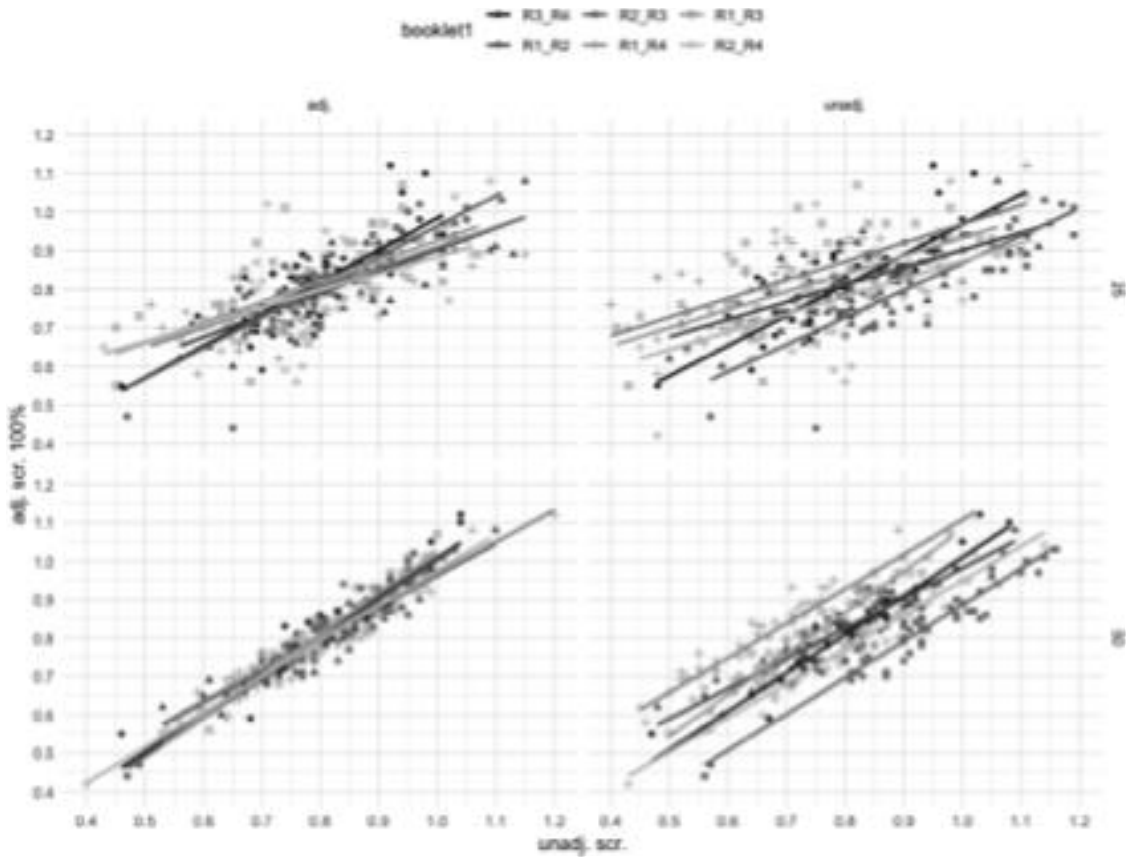


Figure 2. Comparing benchmark scores (adjusted score with full dataset *adj. scr.* 100%; Y-axis) with scores in four conditions of completeness and adjustment: adjusted 25%-dataset (upper-left), unadjusted 25%-dataset (upper-right), adjusted 50%-dataset (lower-left), unadjusted 50%-dataset (lower-right). The different shades and shape of the points indicate the six possible pairwise combinations of raters scoring the ideas (the six booklets). Six regression lines predict benchmark scores from incomplete data conditioned on particular pairwise combinations of raters.

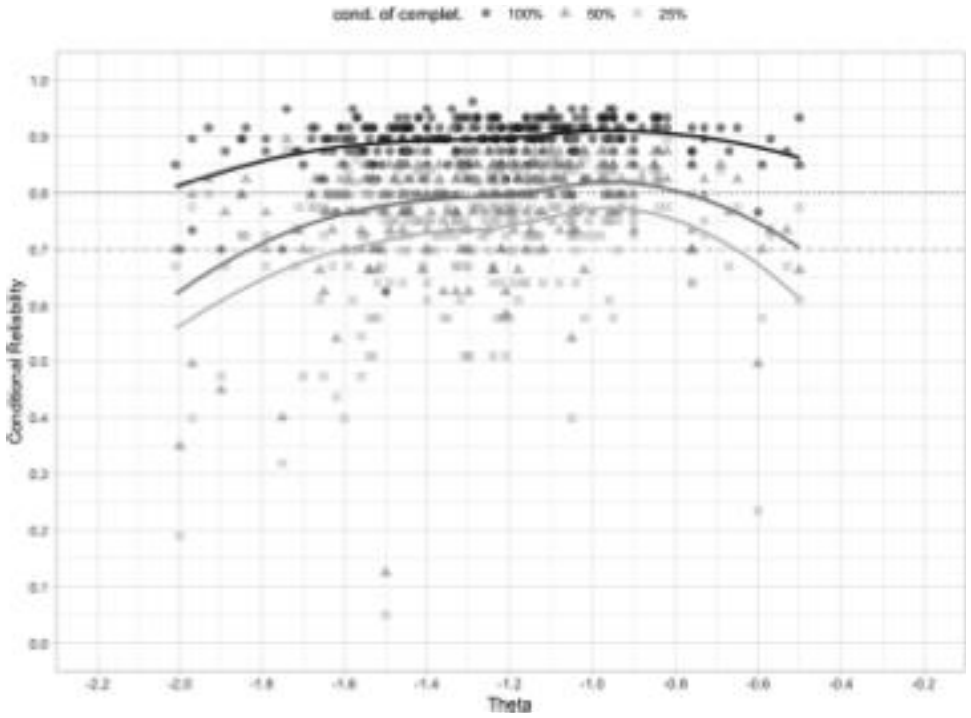


Figure 3. Conditional reliability of latent scores (theta) under three conditions of completeness (black: full data-set, dark-gray: 50%-dataset and light-gray: 25%-dataset). We rescaled results of the information function to the standard reliability metric of 0 to 1, and plotted smoothed average lines of points (i.e., subjects) in these three conditions.

Various Conditions of Completeness and Adjustment With Criterion Measures

unadj. scr. (100%)	*2. adj. scr. (100%)*	*3. unadj. scr. (50%)*	*4. adj. scr. (50%)*	*5. unadj. scr*
.06	.07	.05	.06	.08
.34**	.36**	.24**	.30**	.20**
.16**	.15**	.07	.13*	.03
.22**	.22**	.13*	.18**	.07
.22**	.21**	.13*	.18**	.12*
.37**	.37**	.28**	.34**	.22**

ity measures: unadj. scr: average scores not adjusted for rater effects, adj. scr.: average score
unt of data used to calculate scores varied across three levels: full dataset 100%, 50% and 2
uents (Diedrich et al., 2018).
: .001.

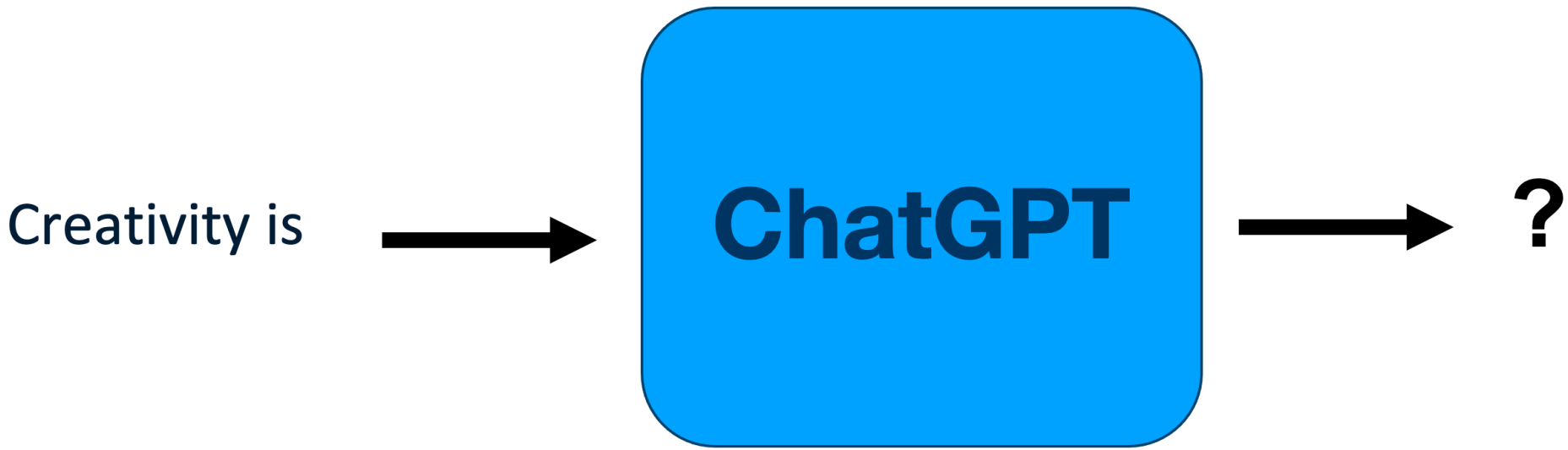
Coding Time:

How to plan a Balanced Incomplete Block Design

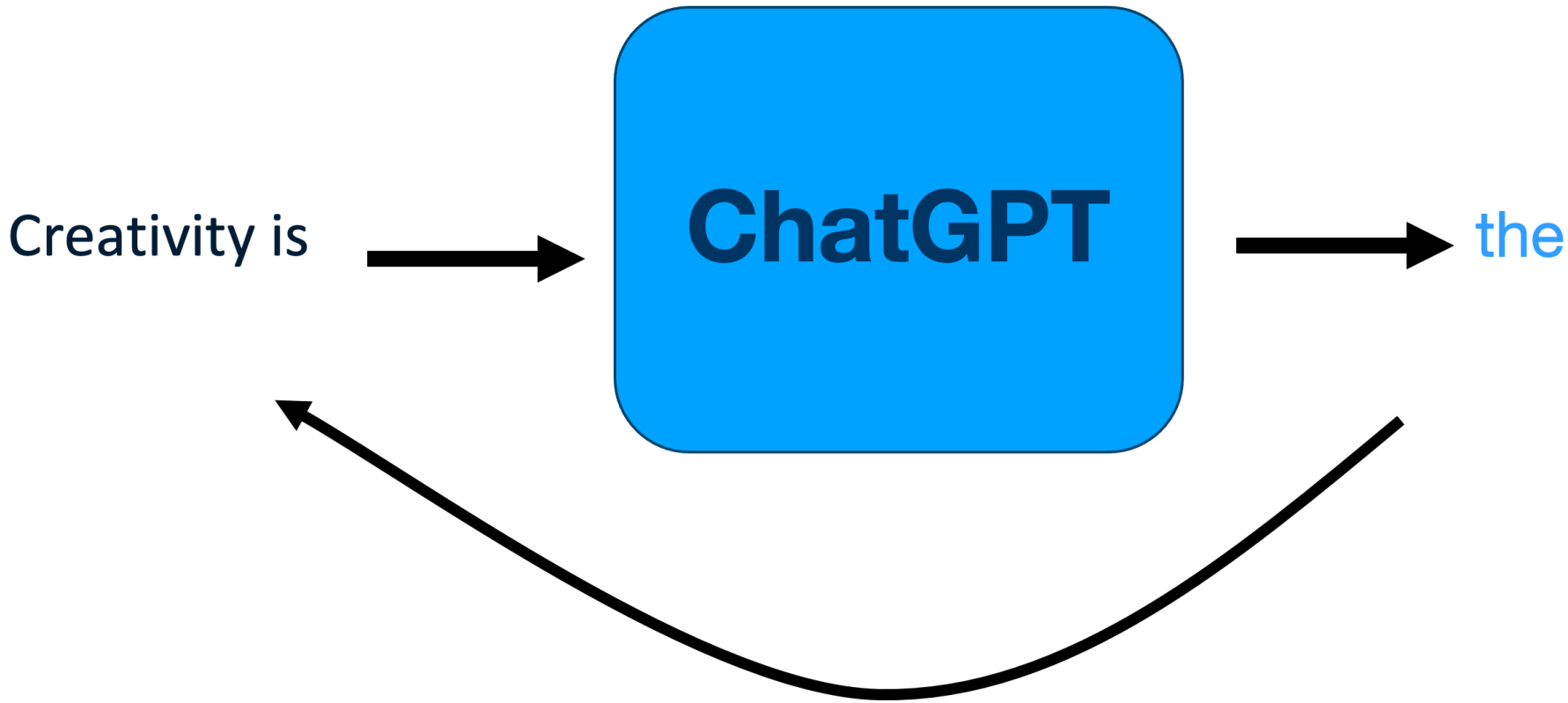
OECD rescoring project (Brazil 8 judges and Portugal 6 judges)

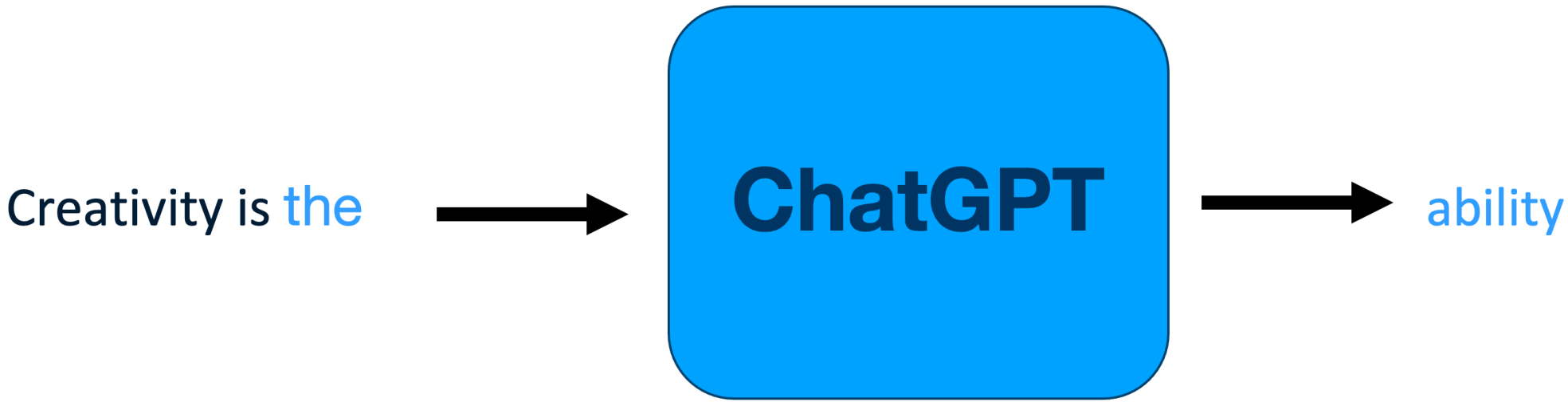
How to analyze data using MFRM

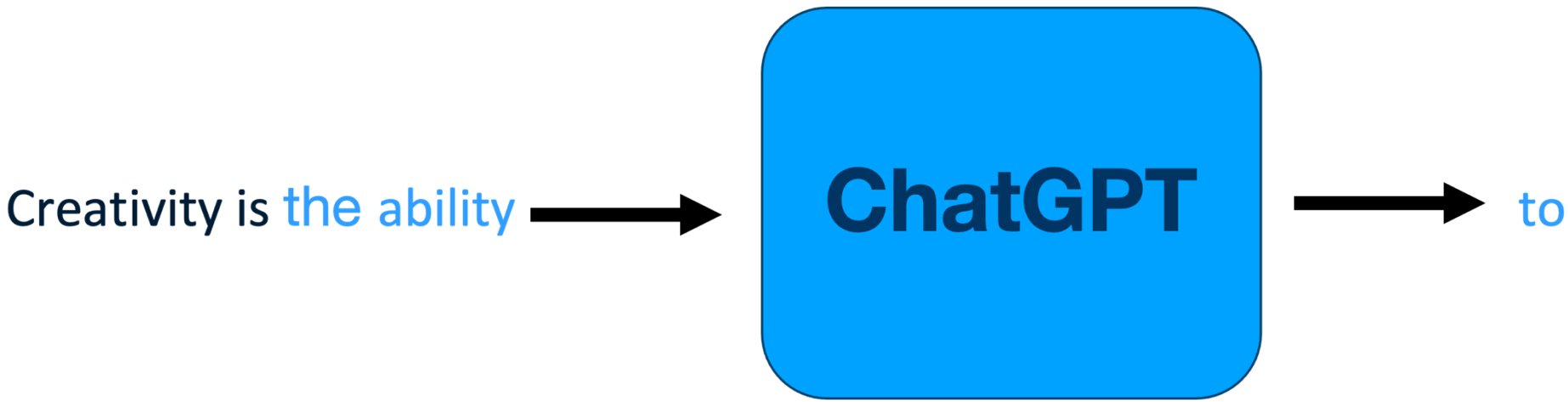
IA and Transformers


Creativity is
ChatGPT
?


Creativity is
ChatGPT
the

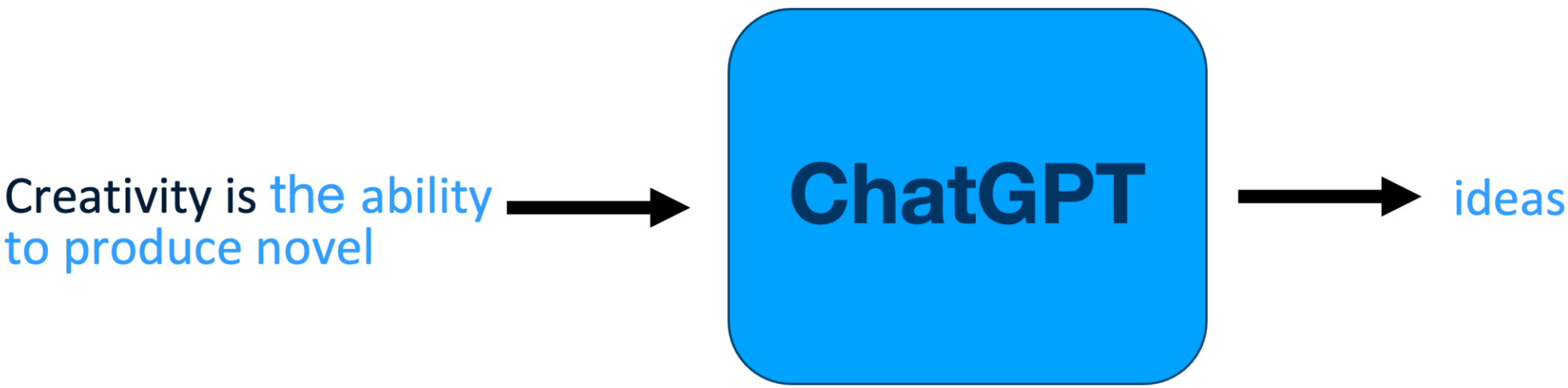

Creativity is
ChatGPT
the


Creativity is the
ChatGPT
ability


Creativity is the ability
ChatGPT
to

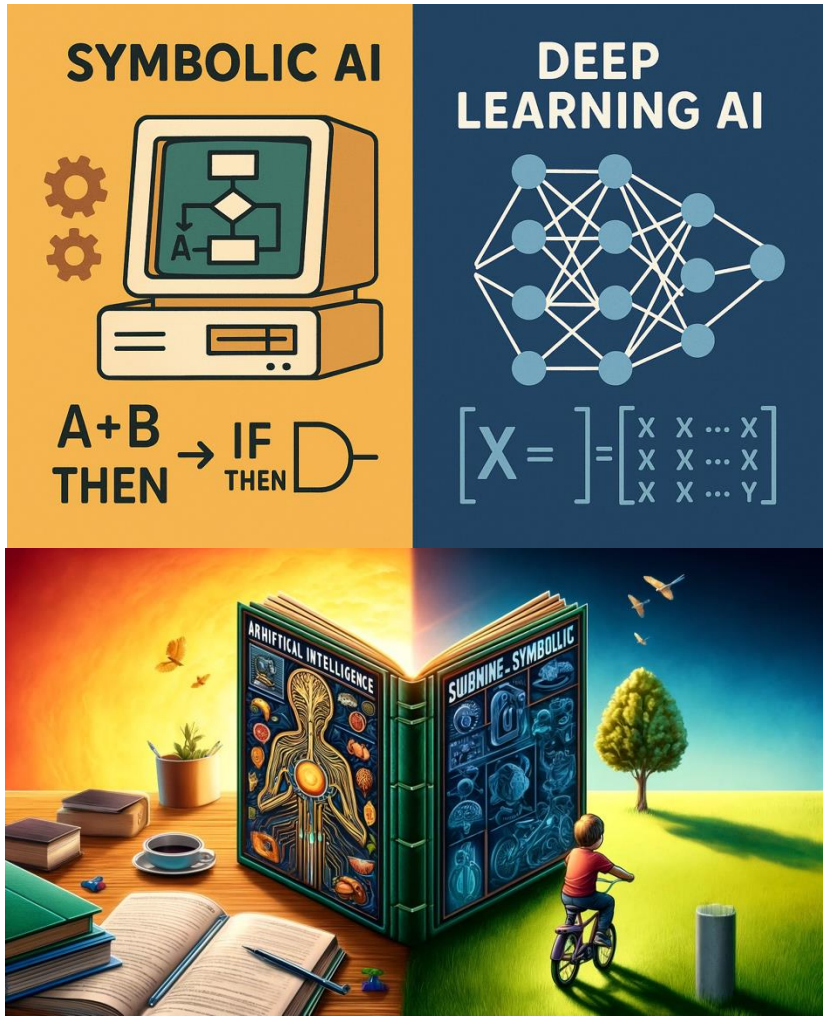
What is generative AI ?

A machine that .... predict next word, complete texts word by word



How does it work? What's 'inside' ChatGPT?

- Deep Neural Network
  - Symbolic (Newell & Simon general problem solver) vs Parallel processing (Rumelhart & McClelland Parallel Distributed Processing)
- Embeddings
  - Latent Semantic Analysis
- Special kind of neural network architecture: attention mechanism
  - Correlation between word vectors
- Language Model
  - $P(wi|w1,w2,\ldots,wi-1)$
- Scaling Laws:
  - Key factors of a big monster: model size, training method, data size, and data quality



How LLMs Learn: A Different Kind of Machine

🛠️ **Symbolic AI (Old Approach – "Build the Car")**

- Human engineers write explicit **rules** and **logic**
- Like assembling a car from blueprints
- Works well for clear, rule-based tasks (e.g., chess, arithmetic)

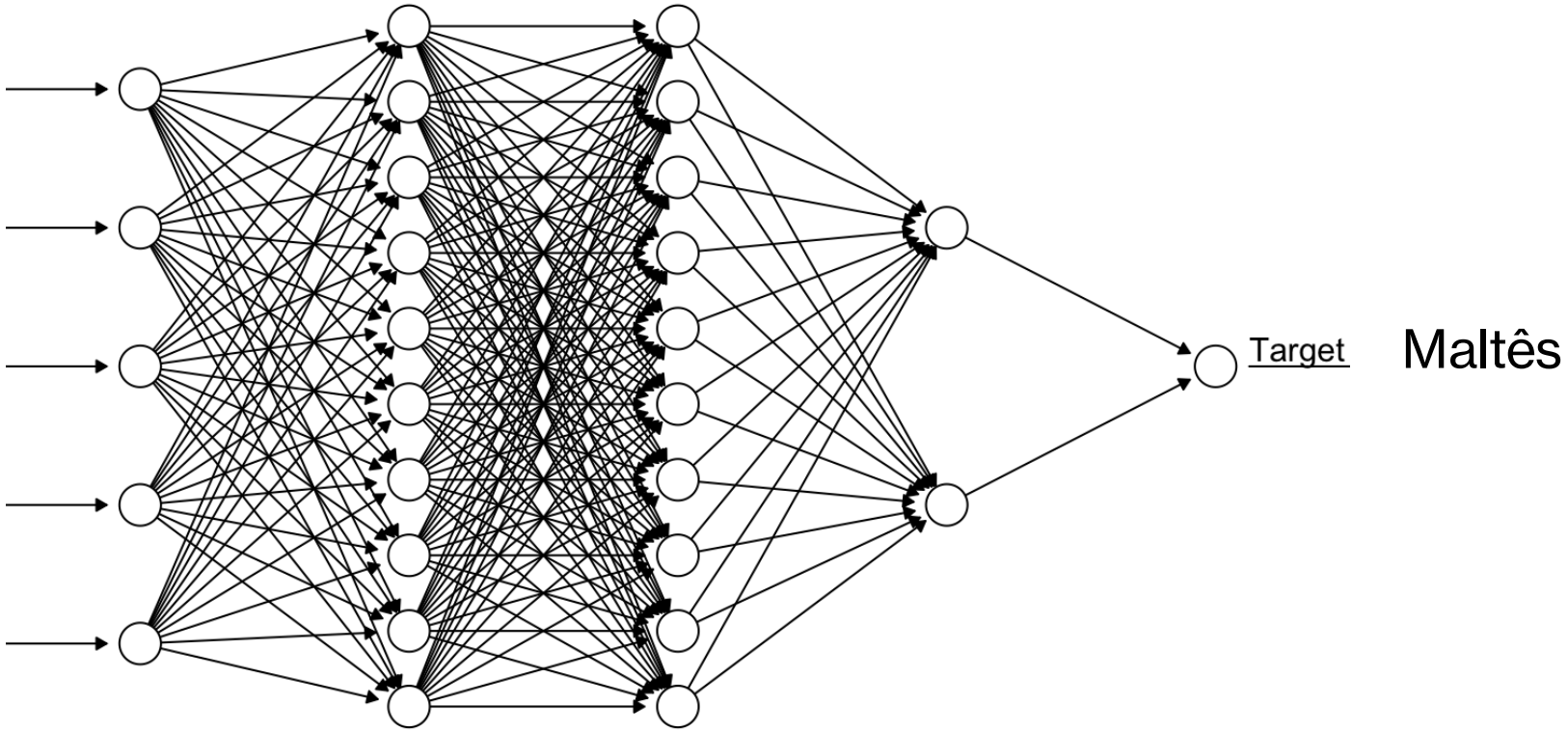
🧠 **LLMs & Deep Learning (New Approach – "Grow the Brain")**

- Build a simplified **neural network** (like a brain made of artificial neurons)
- **Train it** using huge amounts of text (not rules!)
- Goal: *learn statistical patterns* between words and ideas

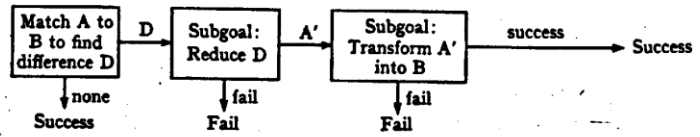
🚀 **How ChatGPT Was Trained**

1. Pre-training: Show it **millions of texts**, with **last word removed**
2. Ask: "What word is likely to come next?"
3. It learns to assign **high probabilities to likely words**
4. Over time, it captures grammar, meaning, abstraction symbolic languages

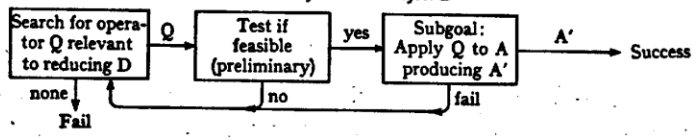
Deep Neural Network



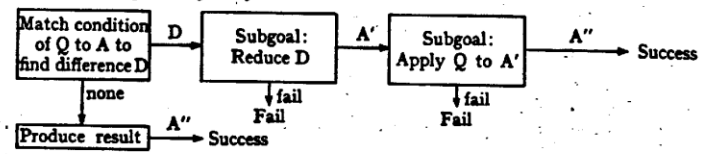
Goal: *Transform object A into object B*



Goal: *Reduct difference D between object A and object B*



Goal: *Apply operator Q to object A*



For the logic task of the text:

Feasibility test (preliminary):

Is the main connective the same? (E. g., A · B → B fails against P ∨ Q)
Is the operator too big? (E. g., (A ∨ B) · (A ∨ C) → A ∨ (B · C) fails against P · Q)
Is the operator too easy? (E. g., A → A · A · A applies to anything)
Are the side conditions satisfied? (E. g., *R8* applies only to main expressions)

Table of connections

	R1	*R2*	*R3*	*R4*	*R5*	*R6*	*R7*	*R8*	*R9*	*R10*	*R11*	*R12*
Add terms			•				•		•	•	•	•
Delete terms			•				•	•			•	•
Change connective					•	•	•					
Change sign					•							
Change lower sign		•			•	•						
Change grouping							•					
Change position	•	•										

• means some variant of the rule is relevant. GPS will pick the appropriate variant.

Figure 3. Methods for GPS

A HEURISTIC PROGRAM TO SOLVE GEOMETRIC-ANALOGY PROBLEMS

Thomas G. Evans
Air Force Cambridge Research Laboratories (OAR)
Bedford, Massachusetts

INTRODUCTION

The purpose of this paper is to describe a program now in existence which is capable of solving a wide class of the so-called 'geometric-analogy' problems frequently encountered on intelligence tests. Each member of this class of problems consists of a set of labeled line drawings. The task to be performed can be concisely described by the question: 'figure A is to figure B as figure C is to which of the given answer figures?' For example, given the problem illustrated as Fig. 1, the geometric-analogy program (which we shall subsequently call ANALOGY, for brevity) selected the problem figure labeled 4 as its answer. It seems safe to say that most people would agree with ANALOGY's answer to this problem (which, incidentally, is taken from the 1942 edition of the *Psychological Test for College Freshmen* of the American Council on Education). Furthermore, if one were required to make explicit the reasoning by which he arrived at his answer, prospects are good that the results would correspond closely to the description of its 'reasoning' produced by ANALOGY.

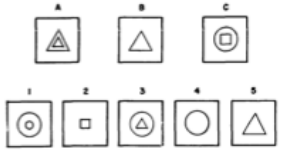


Figure 1.

At this point, a large number of questions might reasonably be asked by the reader. Four, in particular, are:

(i) Why were problems of this type chosen as subject matter?

(ii) How does ANALOGY go about solving these problems?

(iii) How competent is ANALOGY at its subject matter, especially in comparison to human performance?

(iv) What has been learned in the construction of ANALOGY and what implications might this study have for the further development of problem-solving programs in general?

The remainder of this paper constitutes an attempt to answer these questions in some detail. We first deal with a variety of motivations for this investigation and attempt to place it in the context of other work in related areas. Next we turn to detailed consideration of the problem type and of the mechanism of the ANALOGY program. Finally, we present some answers to

Evans, T. G. (1968). Program for the solution of a class of geometric-analogy intelligent-test questions. In: M. Minsky (Org), *Semantic information processing*. Cambridge, MA: MIT Press.

PARALLEL DISTRIBUTED PROCESSING

Explorations in the Microstructure of Cognition

Volume 1: Foundations

David E. Rumelhart James L. McClelland
and the PDP Research Group

Chisato Asanuma
Francis H. C. Crick
Jeffrey L. Elman
Geoffrey E. Hinton
Michael I. Jordan
Alan H. Kawamoto
Paul W. Munro
Donald A. Norman
Daniel E. Rabin
Terrence J. Sejnowski
Paul Smolensky
Gregory O. Stone
Ronald J. Williams
David Zipser

Institute for Cognitive Science
University of California, San Diego

A Bradford Book

The MIT Press
Cambridge, Massachusetts
London, England

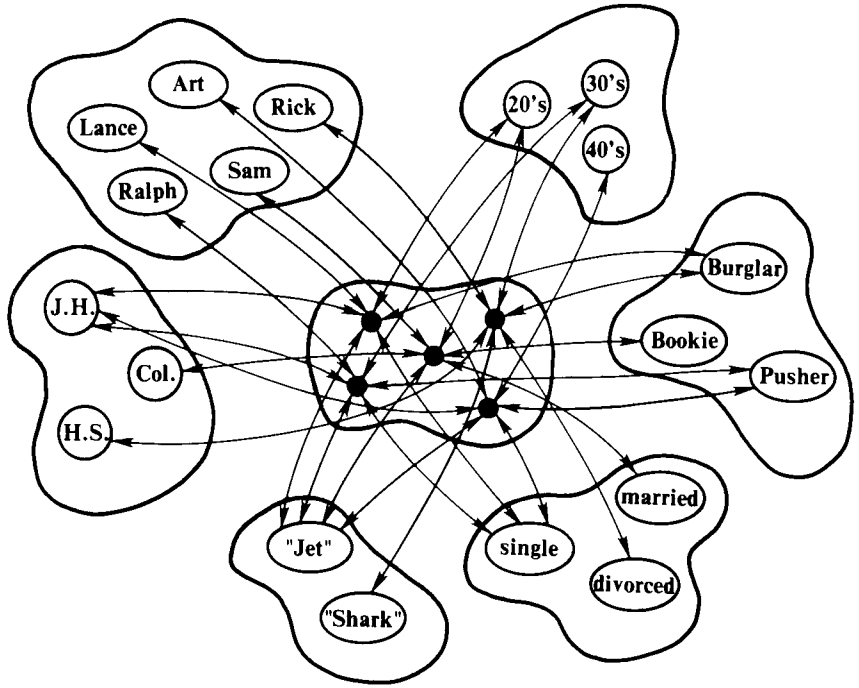


FIGURE 11. Some of the units and interconnections needed to represent the individuals shown in Figure 10. The units connected with double-headed arrows are mutually excitatory. All the units within the same cloud are mutually inhibitory. (From "Retrieving General and Specific Knowledge From Stored Knowledge of Specifics" by J. L. McClelland, 1981, *Proceedings of the Third Annual Conference of the Cognitive Science Society*, Berkeley, CA. Copyright 1981 by J. L. McClelland. Reprinted by permission.)

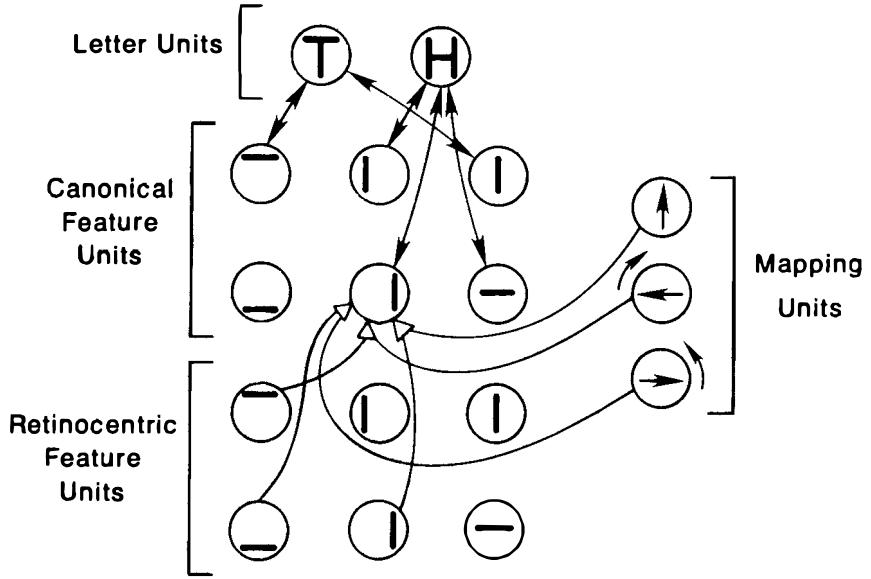


FIGURE 2. Hinton's (1981b) scheme for mapping patterns in one coordinate system into patterns in another coordinate system. At the top are two letter-detector units, with mutual excitatory connections to the six canonical feature units (the position and orientation of the line segment each of these detectors represents is indicated by the line segment in the "body" of each unit). At the bottom are six retinocentric feature units, and at the right are units corresponding to each of three different mappings from retinocentric to canonical features. (The arrows on the units indicate which direction in the retinocentric frame corresponds to upright in the canonical frame, and the arrow outside the unit indicates the nature of the transformation imposed on the retinocentric pattern). Each canonical unit receives three pairs of inputs, with each pair arriving at a multiplicative connection. These inputs are illustrated for one canonical unit only.

PARALLEL DISTRIBUTED PROCESSING

Explorations in the Microstructure of Cognition

Volume 1: Foundations

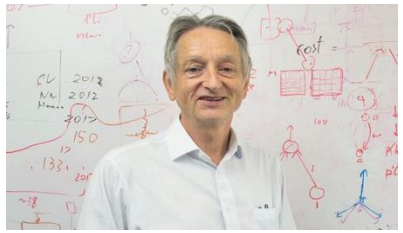
David E. Rumelhart James L. McClelland
and the PDP Research Group

Chisato Asanuma, Francis H. C. Crick, Jeffrey L. Elman, Geoffrey E. Hinton, Michael I. Jordan, Alan H. Kawamoto, Paul W. Munro, Donald A. Norman, Daniel E. Rabin, Terrence J. Sejnowski, Paul Smolensky, Gregory O. Stone, Ronald J. Williams, David Zipser

Institute for Cognitive Science
University of California, San Diego

A Bradford Book

The MIT Press
Cambridge, Massachusetts
London, England



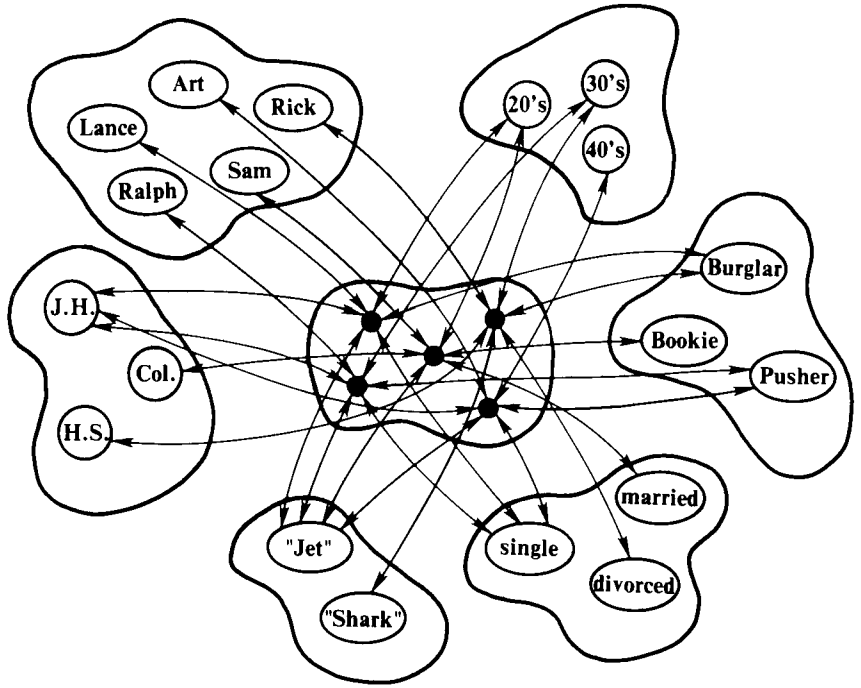


FIGURE 11. Some of the units and interconnections needed to represent the individuals shown in Figure 10. The units connected with double-headed arrows are mutually excitatory. All the units within the same cloud are mutually inhibitory. (From "Retrieving General and Specific Knowledge From Stored Knowledge of Specifics" by J. L. McClelland, 1981, *Proceedings of the Third Annual Conference of the Cognitive Science Society*, Berkeley, CA. Copyright 1981 by J. L. McClelland. Reprinted by permission.)

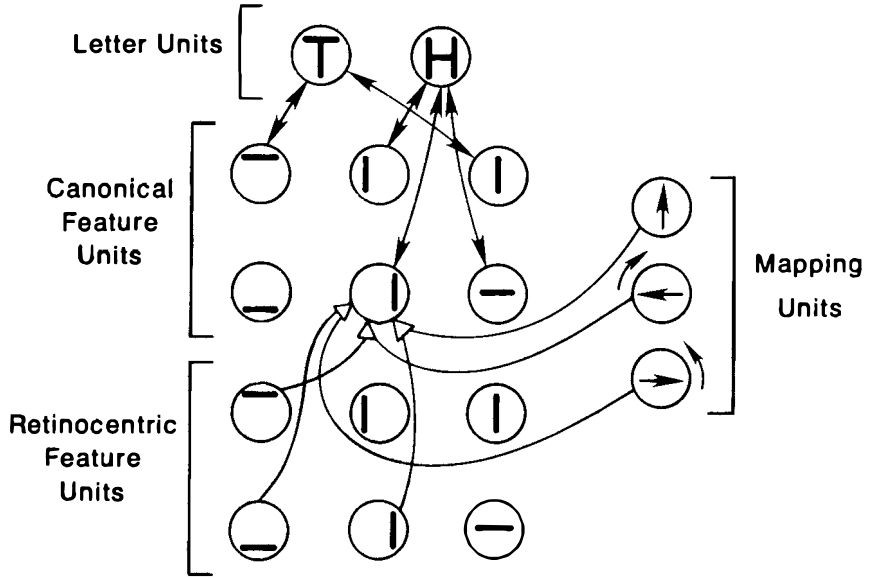
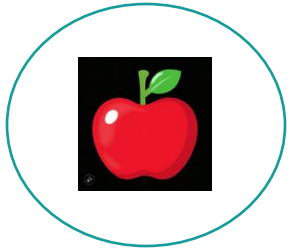
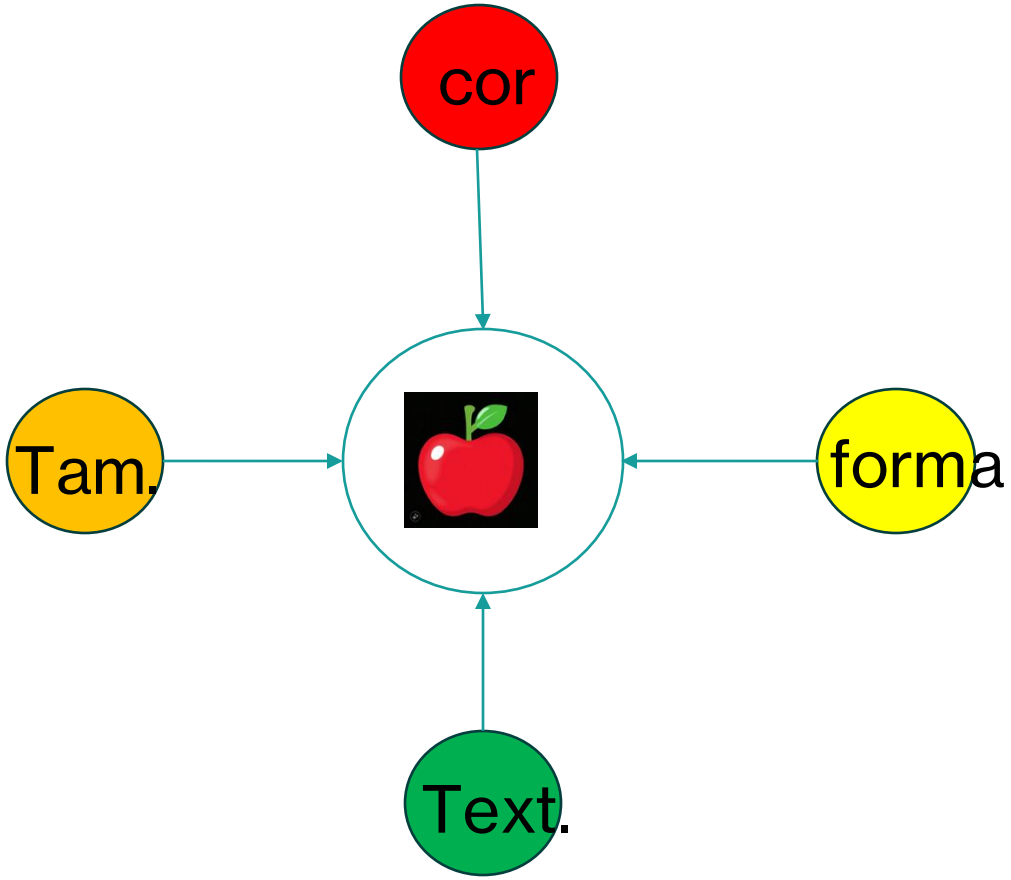


FIGURE 2. Hinton's (1981b) scheme for mapping patterns in one coordinate system into patterns in another coordinate system. At the top are two letter-detector units, with mutual excitatory connections to the six canonical feature units (the position and orientation of the line segment each of these detectors represents is indicated by the line segment in the "body" of each unit). At the bottom are six retinocentric feature units, and at the right are units corresponding to each of three different mappings from retinocentric to canonical features. (The arrows on the units indicate which direction in the retinocentric frame corresponds to upright in the canonical frame, and the arrow outside the unit indicates the nature of the transformation imposed on the retinocentric pattern). Each canonical unit receives three pairs of inputs, with each pair arriving at a multiplicative connection. These inputs are illustrated for one canonical unit only.




cor
Tam.
forma
Text.

COGNITIVE PSYCHOLOGY 5, 1–28 (1973)

A Model for Analogical Reasoning[1]

DAVID E. RUMELHART AND ADELE A. ABRAHAMSON
University of California, San Diego

A theory of analogical reasoning is proposed in which the elements of a set of concepts, e.g., animals, are represented as points in a multidimensional Euclidean space. Four elements A,B,C,D, are in an analogical relationship A:B::C:D if the vector distance from A to B is the same as that from C to D. Given three elements A,B,C, an ideal solution point I for A:B::C:? exists. In a problem A:B::C:D_1, . . ., D_i, . . ., D_n, the probability of choosing D_i as the best solution is a monotonic decreasing function of the absolute distance of D_i from I. A stronger decision rule incorporating a negative exponential function in Luce's choice rule is also proposed. Both the strong and weak versions of the theory were supported in two experiments where Ss rank-ordered the alternatives in problems A:B::C:D_1,D_2, $D_3$$D_4$. In a third experiment the theory was applied and further tested in teaching new concepts by analogy.

To introduce our notion of analogical reasoning, it is useful to outline a definition of the word *reasoning* from which we can work. The term is used here to denote those processes in information retrieval which depend on the *structure*, as opposed to the *content* of organized memory.

Thus, one might answer the question "Who is the father of your country?" in at least two different ways. In one case, the specific information that George Washington was the "father of our country" might be stored and used to answer the question. On the other hand, when specific information is not available, one can consult the stored meanings of the words in question and one's knowledge of history to derive a plausible answer. The first of these methods might be called remembering, since retrieval depends on the specific information stored. The second method may be identified with reasoning, since in this case retrieval depends to a much greater extent on the *form* of the relationship among the words. The same act of reasoning (i.e., the same processes) could have been applied to the question "Who was the father of your state?" or "Who was the mother of your country?". It is not the specific content of the question but the form of the relationships among the words which determines the response.

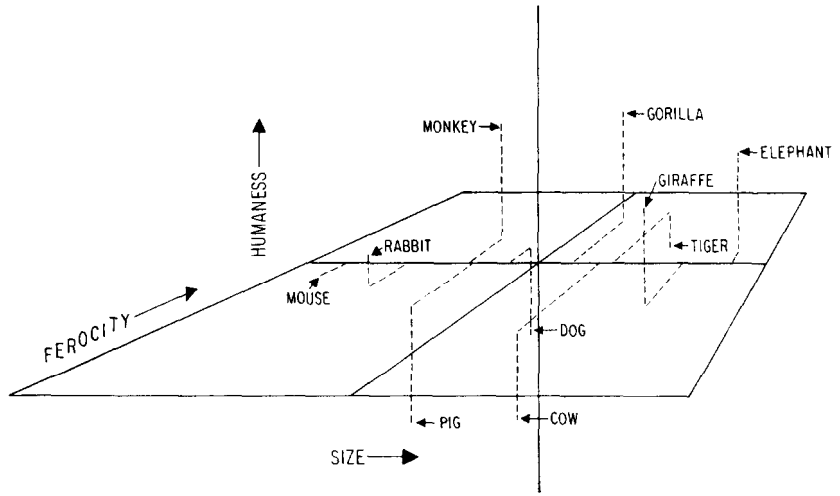


FIG. 1. The placements of a selected set of animals based on data from Henley (1969).

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

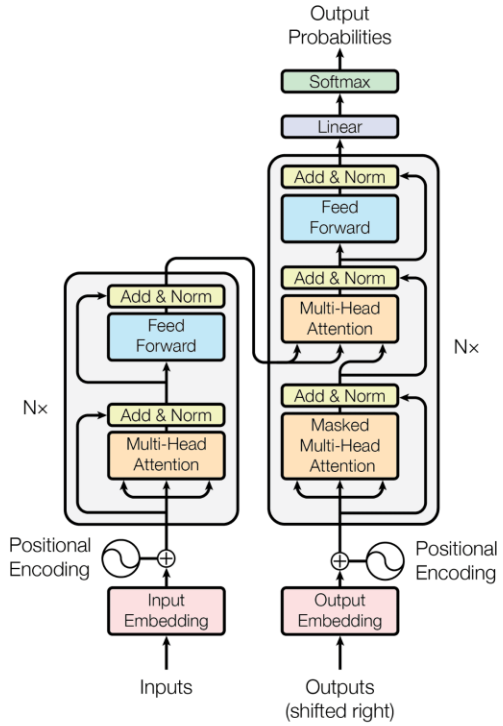
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

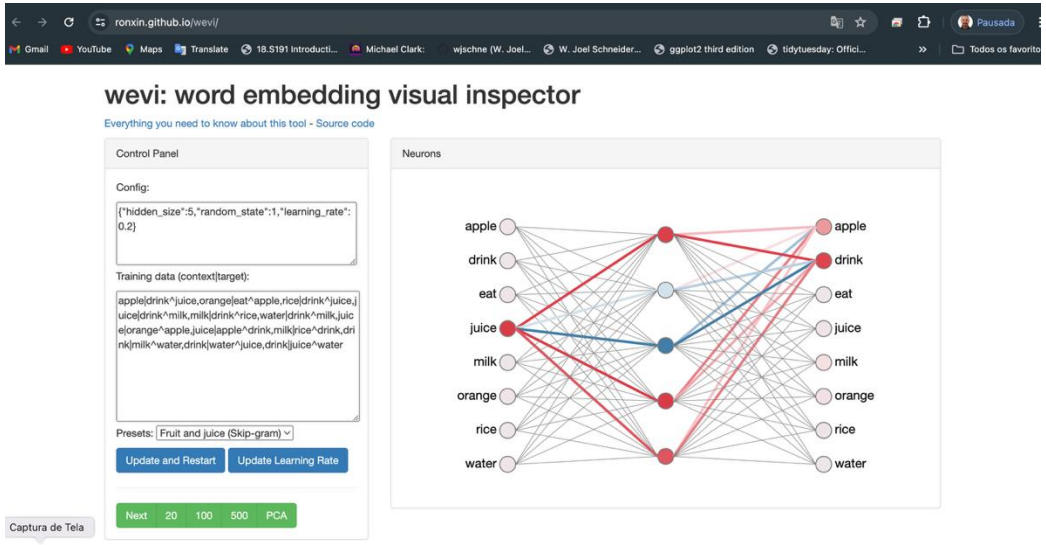
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

GPT: **G**enerative **P**re-trained **T**ransformer.
BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
ChatGPT: GPT post trained with Reinfocment Learning from Human Feedback

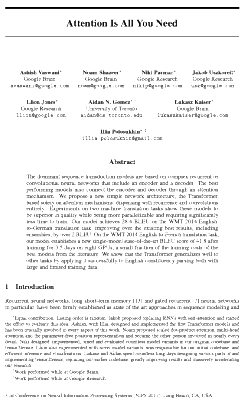


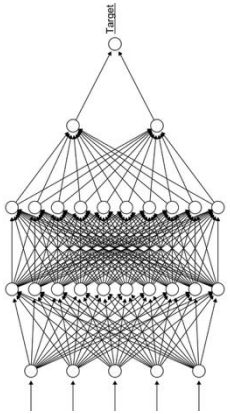
https://www.3blue1brown.com/topics/neural-networks
chap 5 and 6

https://ronxin.github.io/wevi/



Simplified scheme of a Transformer





N layers

P (vocabulário)

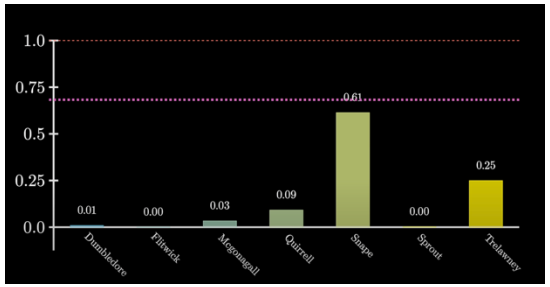
Output

Traditional Neural Network

Attention Layer

Embeddings

Input Tokens



Produces a distribution of the probabilities for the words in the vocabulary being the next word

Transformed word vectors representing the meaning of words/text

Calculates interaction between word vectors (relationship between word semantic meanings).

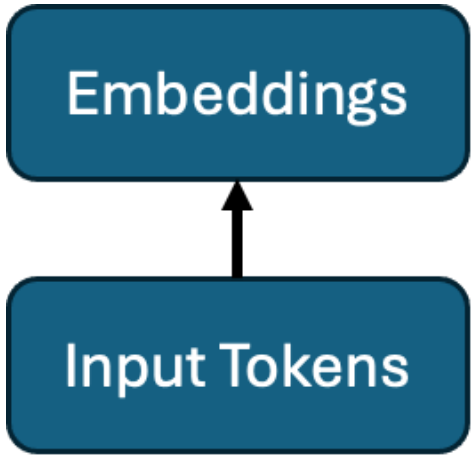
Transforms words into numerical vectors (semantic meaning)

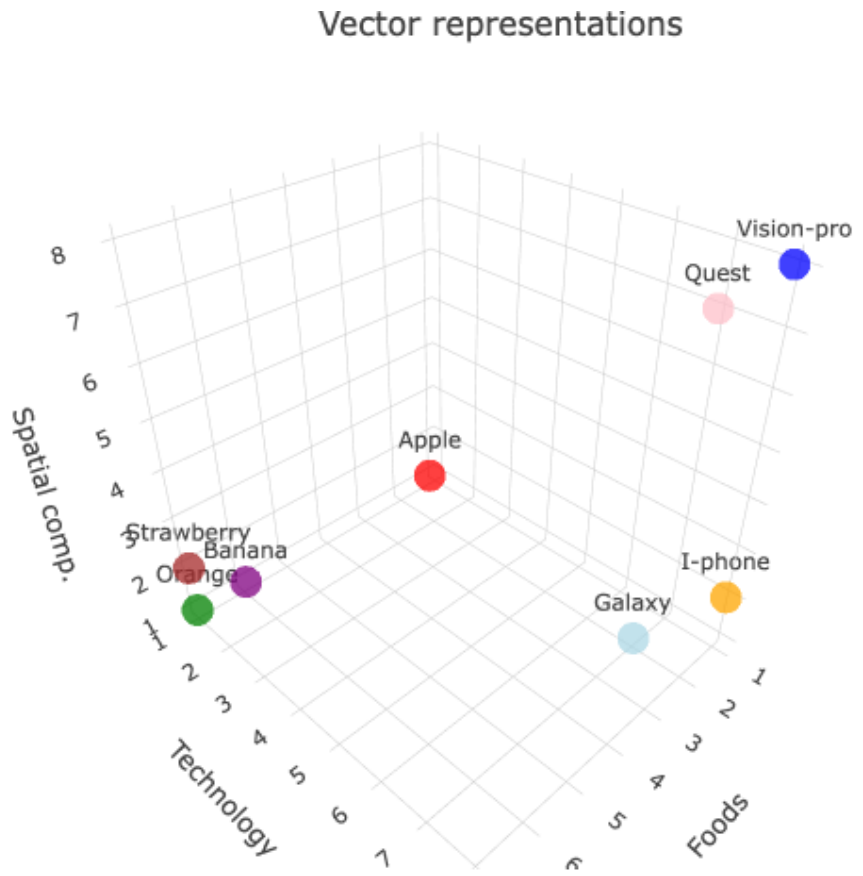
Embeddings

From Words to Meaning: How LLMs Represent Language

- Words (tokens) are converted into **vectors of continuous numbers** → *embeddings*
- Embeddings capture **relationships** between words
- Encode **semantic meaning**, not just surface form
- Encode **deep sbstractions structures**: grammar, code, symbolic math, logic systems
  - Emerged from training on **billions of words** from diverse sources and Large context windows that let models analyze **meaning across long passages**

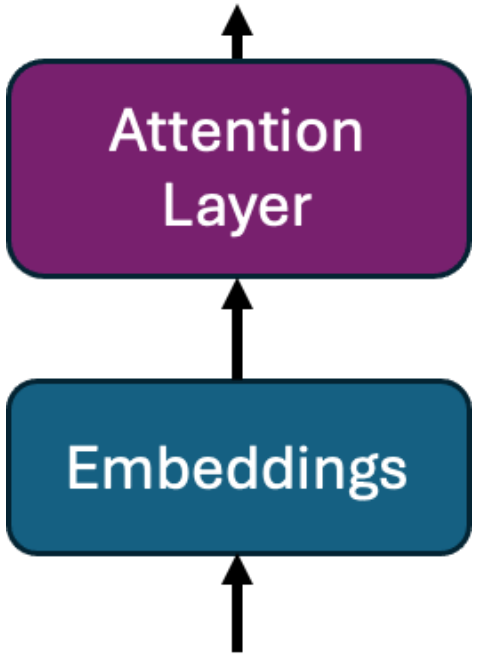
I bought a box of apples.
I bought an apple i-phone.



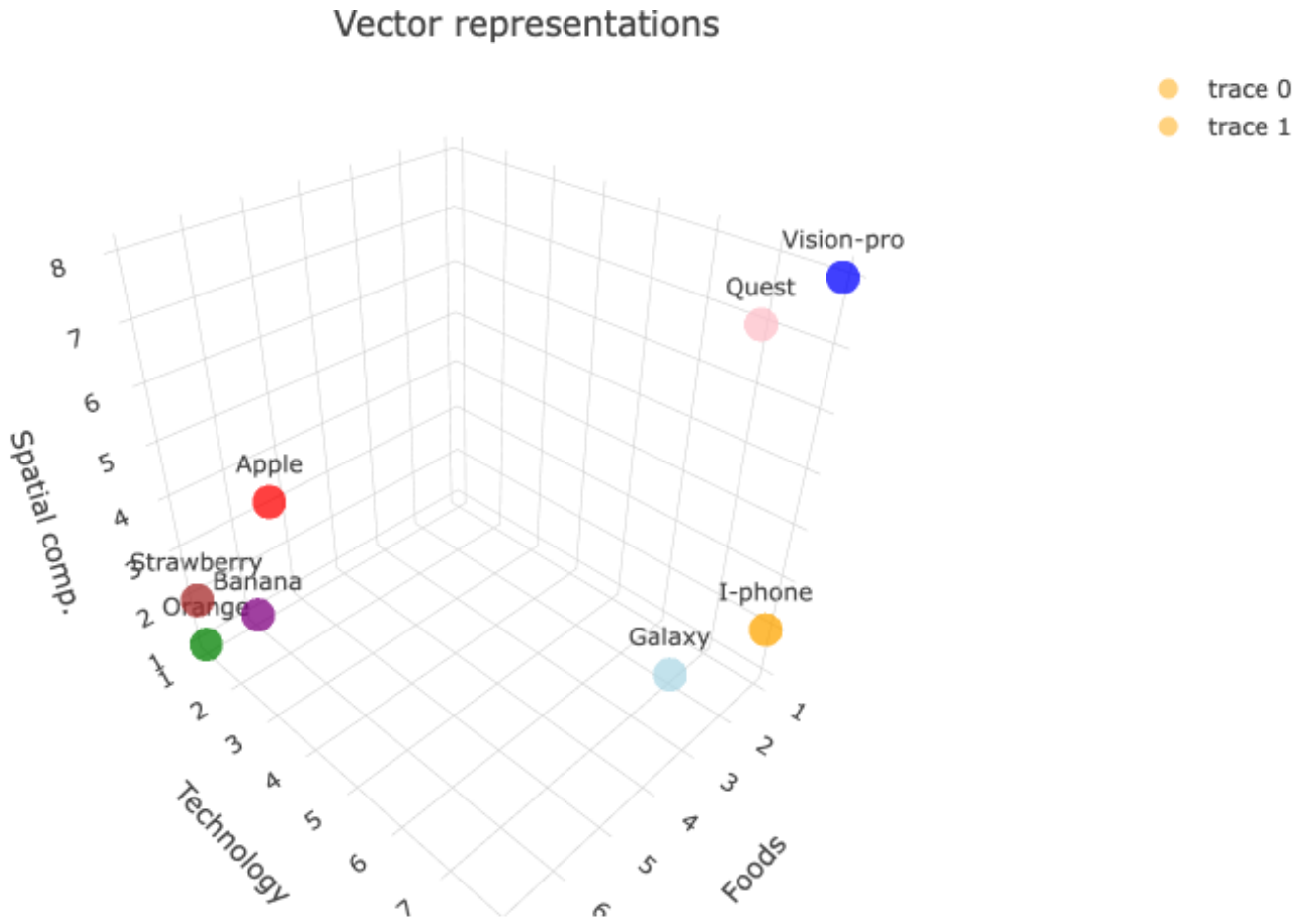


	objs	tech	alim	spatial	colors
1	I-phone	8.1	1.1	2.1	orange
2	Galaxy	7.0	2.0	1.0	lightblue
3	Vision-pro	8.0	1.0	8.0	blue
4	Orange	1.0	7.0	1.0	green
5	Banana	2.0	6.5	2.0	purple
6	Apple	4.0	4.0	4.0	red
7	Strawberry	1.0	7.0	2.0	brown
8	Quest	7.0	1.0	7.0	pink

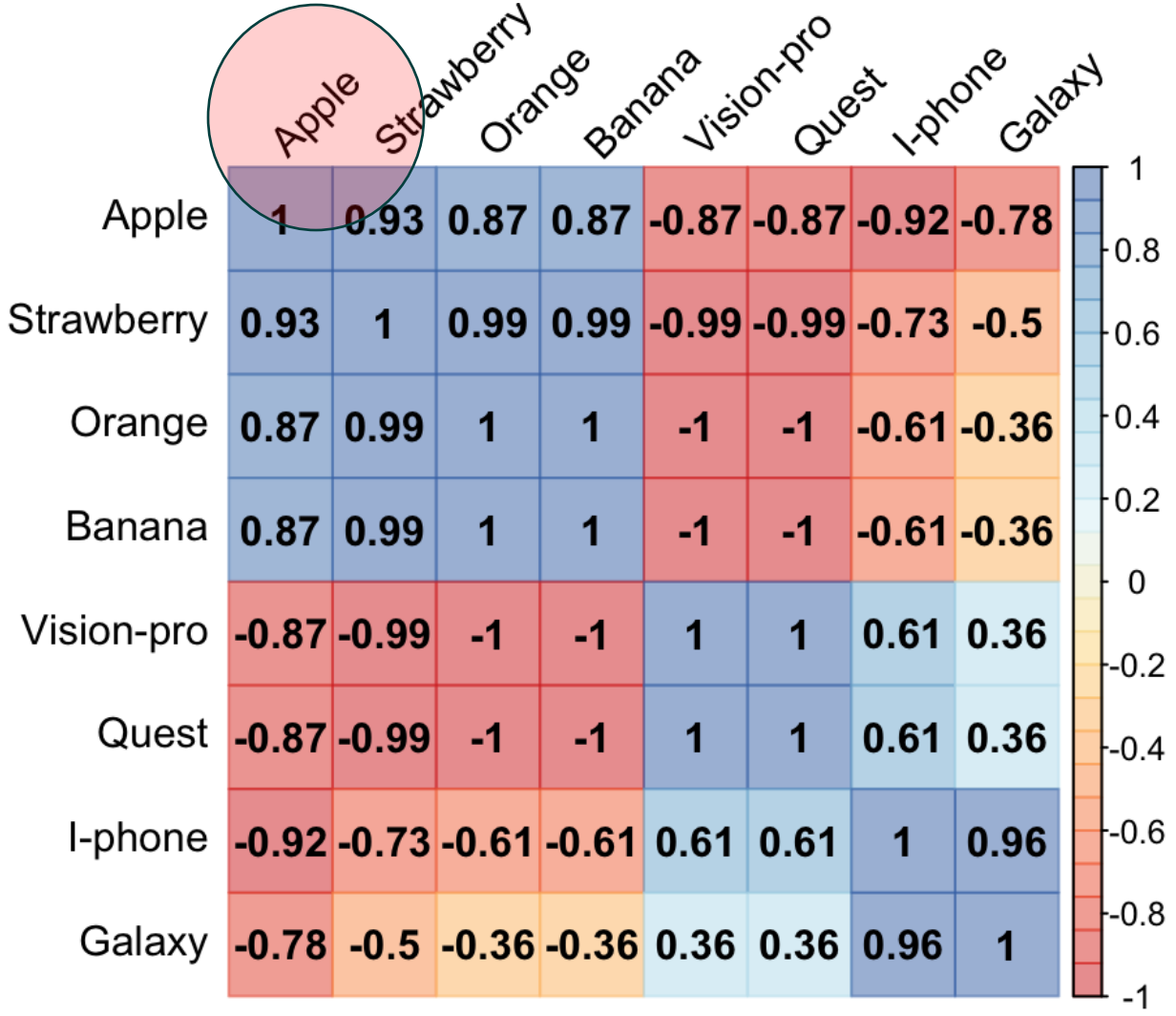
- Each word has a numerical vector of dimension D (GPT-4: 1536 dimensions).
- Numbers are learned such that each dimension encodes a semantic attribute.
- Similar words have similar numerical vectors, which means they are close to each other in the D-dimensional space.
- Therefore, the correlation between vectors (cosine similarity) is higher for similar words.

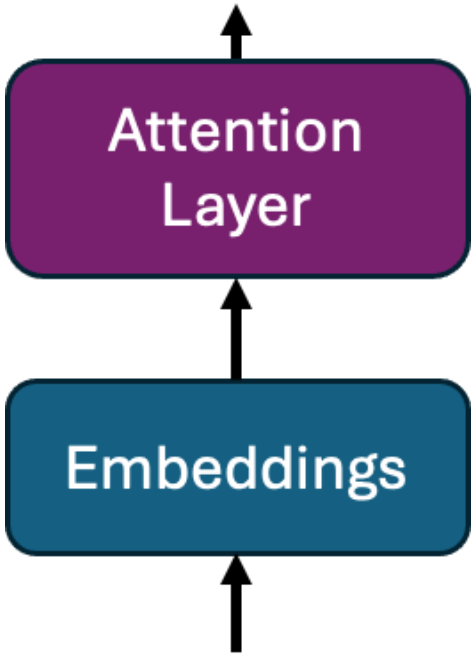


I bought a box of apples

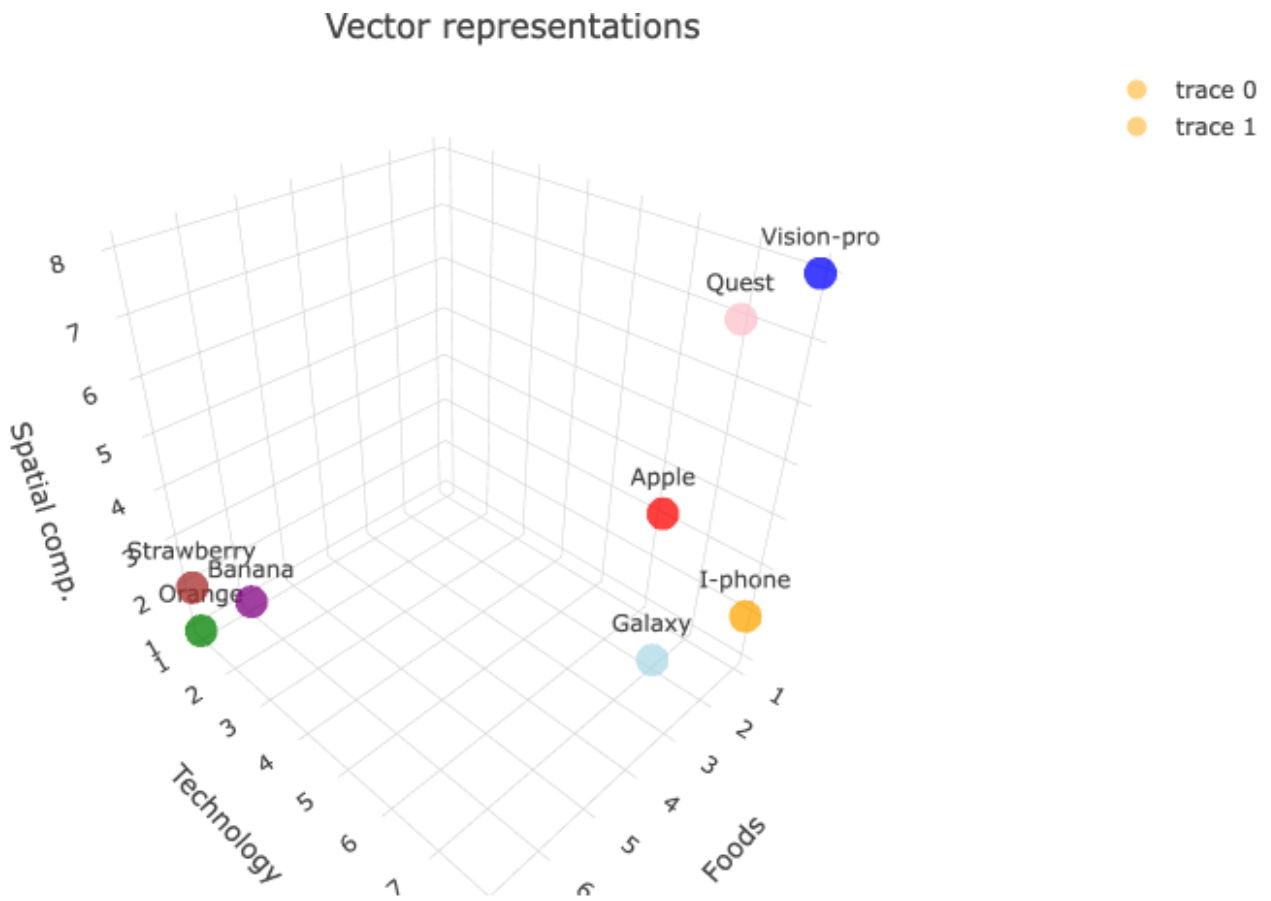


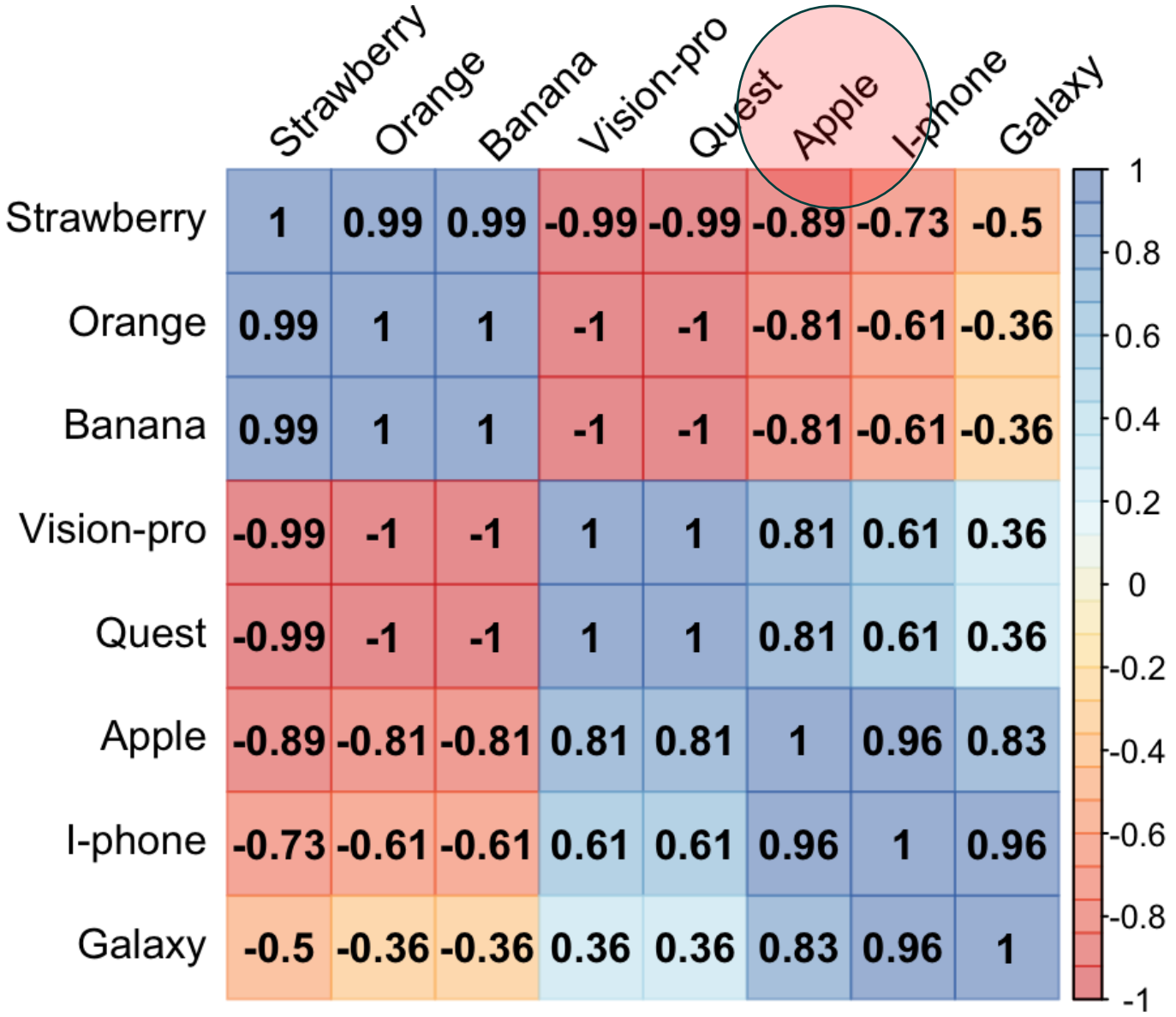
Correlations between embeddings: attention



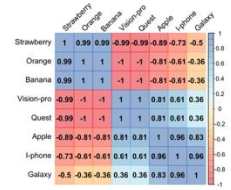


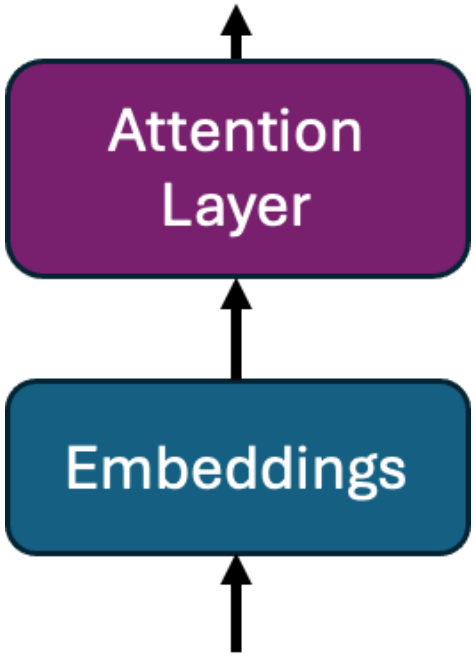
I bought an apple i-phone.




Strawberry
Orange
Banana
Vision-pro
Quest
Apple
I-phone
Galaxy
Strawberry 1 0.99 0.99 -0.99 -0.99 -0.89 -0.73 -0.5
Orange 0.99 1 1 -1 -1 -0.81 -0.61 -0.36
Banana 0.99 1 1 -1 -1 -0.81 -0.61 -0.36
Vision-pro -0.99 -1 -1 1 1 0.81 0.61 0.36
Quest -0.99 -1 -1 1 1 0.81 0.61 0.36
Apple -0.89 -0.81 -0.81 0.81 0.81 1 0.96 0.83
I-phone -0.73 -0.61 -0.61 0.61 0.61 0.96 1 0.96
Galaxy -0.5 -0.36 -0.36 0.36 0.36 0.83 0.96 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Attention matrix





Each word **“attends to” all others** to refine its meaning
Attention layers **capture relationships** across the sentence (not just nearby words)
Embeddings become **context-aware**, enriched with deeper meaning
Context includes not just text – but **task, intent, and document structure**
Longer context windows = better understanding (128k tokens to 2 Milion Gemini, to 10 milion META AI)
Key insight: Transformers are **super-embedders** – they build rich, semantic maps of language

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* † illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

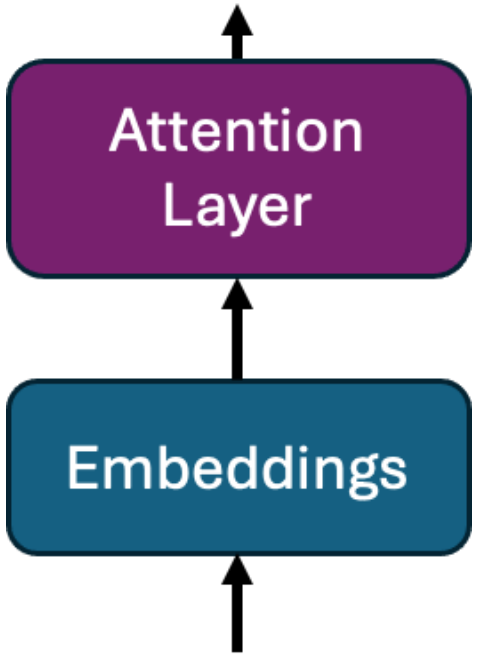
Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

* Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

† Work performed while at Google Brain.

‡ Work performed while at Google Research.

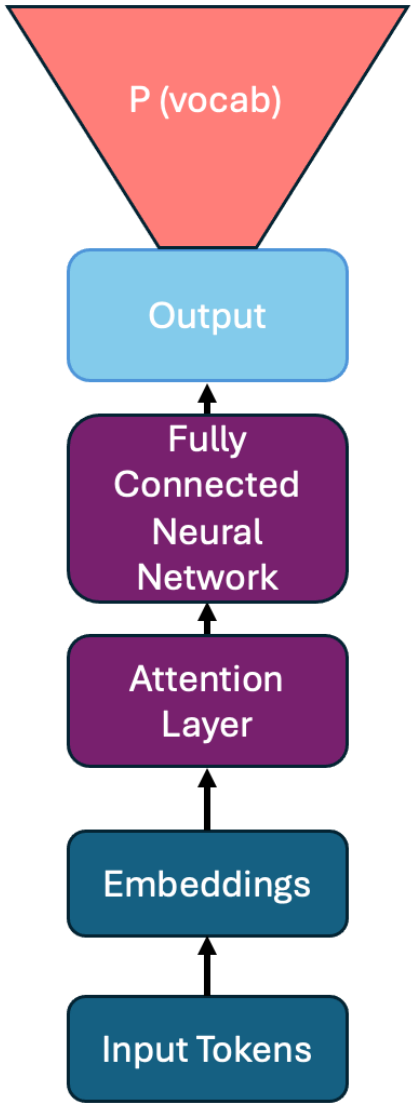
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



Attention matrix

- In the attention layer, the correlations between embeddings are learned!
- These correlations become weights for each word in the sentence to recompose the word vectors.
- For instance, the word vector of "apple" is a weighted sum of all other word vectors, with the weights determined by these correlations.
- Each word "attends" to all the words in the sentence to recompose its meaning.
- The correlation between word vectors serves as the weights in the attention mechanism.

	Strawberry	Orange	Banana	Vision-pro	Quest	Apple	I-phone	Galaxy
Strawberry	1	0.99	0.99	-0.99	-0.99	-0.89	-0.73	-0.5
Orange	0.99	1	1	-1	-1	-0.81	-0.61	-0.36
Banana	0.99	1	1	-1	-1	-0.81	-0.61	-0.36
Vision-pro	-0.99	-1	-1	1	1	0.81	0.61	0.36
Quest	-0.99	-1	-1	1	1	0.81	0.61	0.36
Apple	-0.89	-0.81	-0.81	0.81	0.81	1	0.96	0.83
I-phone	-0.73	-0.61	-0.61	0.61	0.61	0.96	1	0.96
Galaxy	-0.5	-0.36	-0.36	0.36	0.36	0.83	0.96	1



After the attention layer, embeddings obtain richer representations of semantic information from the context.

Contextual Information: The context is not only the information shared among adjacent words but also meta-knowledge about the purpose and task related to a sentence and the high-level meaning of the entire document within its context window.

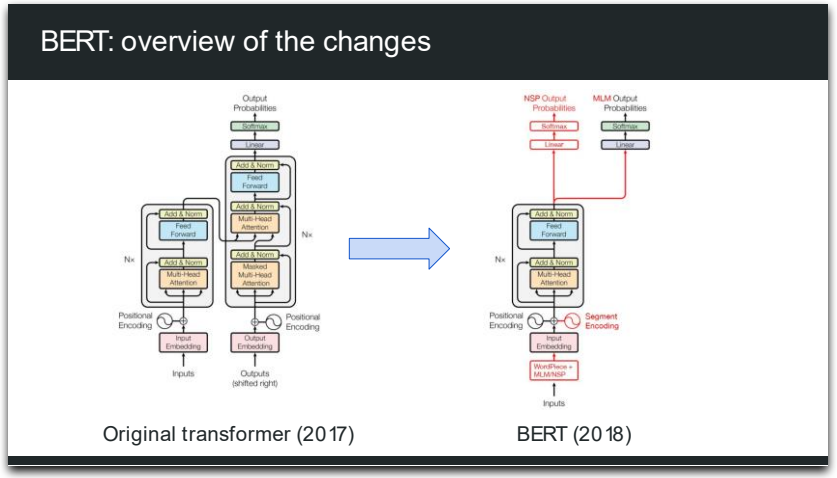
Context Length: The length of the context is crucial to this mechanism:

- GPT-1: 512 tokens
- GPT-4: 128,000 tokens
- Gemini: 2,000,000 tokens

Key Takeaway: The transformer acts as a super embedder, creating highly enriched semantic and contextual representations of each token.

Training methods of a language model and scaling laws

Masked Language Model (MLM)



Training method of BERT
BERT: Bidirectional Encoder Representations from Transformers

"The quick brown fox jumps over the lazy dog."

Cloze Test Sentence:

"The quick brown [MASK] jumps over the lazy [MASK]."

In this example, BERT would be tasked with predicting the masked words based on the context provided by the rest of the sentence. The correct predictions would be "fox" for the first mask and "dog" for the second mask.

Causal Language Modeling
Training method of GPT (Generative Pre-trained Transformer)

Original Sentence:

"The quick brown fox jumps over the lazy dog."

Training Process:

The model is trained to predict each word in the sentence one at a time based on the previous words.

The ___

The quick ___

The quick brown ___

The quick brown fox ___

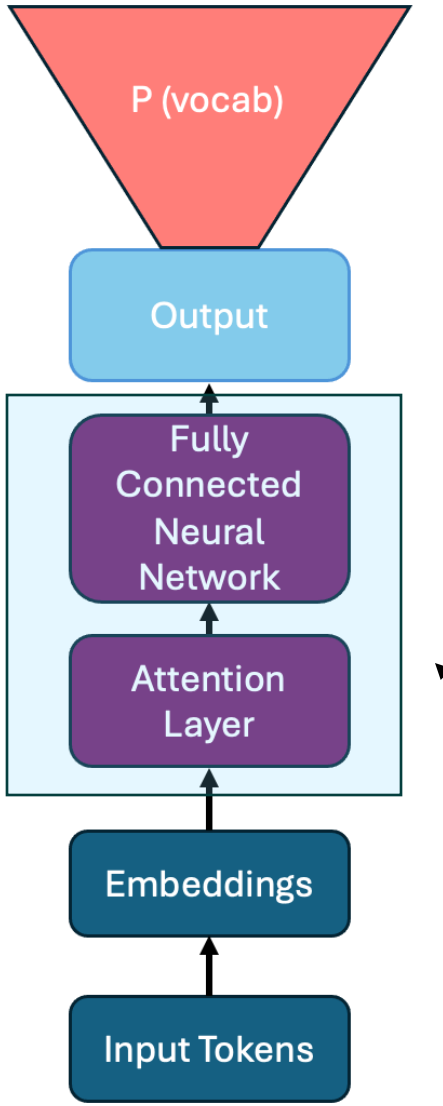
The quick brown fox jumps ___

The quick brown fox jumps over ___

The quick brown fox jumps over the ___

The quick brown fox jumps over the lazy ___

Emergent Abilities and Scaling laws



100 Stacked Transformer Blocks

- Emergent abilities
  - Capabilities that arise when models reach a certain scale, which were not explicitly programmed or expected based on the smaller models.
  - Different from those seen in smaller models
  - Complex understanding and generation tasks that require a deeper comprehension of context, inference, and reasoning
- Scaling laws:
  - Functions that predict emergent abilities from model size, dataset size, compute budget
  - Path to AGI (GPT-5 is trained probably 5X times compute budget than GPT4)

Key Milestones

- GPT-1 (2018): transformer-based architecture with 117 million parameters. Demonstrated the potential of transformers for language modeling but did not exhibit significant emergent abilities.
- GPT-2 (2019): Scaled up to 1.5 billion parameters. Showed improvements in generating coherent text but still had limitations in performing complex tasks.
- GPT-3 (2020): With 175 billion parameters, GPT-3 demonstrated a wide range of emergent abilities, performing tasks such as complex arithmetic, contextual reasoning, and few-shot learning with impressive accuracy.
- GPT-4 (2023) ?1.5 trilion parameters?

Meta-learning or in-context learning

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam

Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss

Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh

Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

Abstract

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

1 Introduction

NLP has shifted from learning task-specific representations and designing task-specific architectures to using task-agnostic pre-training and task-agnostic architectures. This shift has led to substantial progress on many challenging NLP tasks such as reading comprehension, question answering, textual entailment, among others. Even though the architecture and initial representations are now task-agnostic, a final task-specific step remains: fine-tuning on a large dataset of examples to adapt a task agnostic model to perform a desired task.

Recent work [RWC+19] suggested this final step may not be necessary. [RWC+19] demonstrated that a single pretrained language model can be zero-shot transferred to perform standard NLP tasks

*Equal contribution
†Johns Hopkins University, OpenAI

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

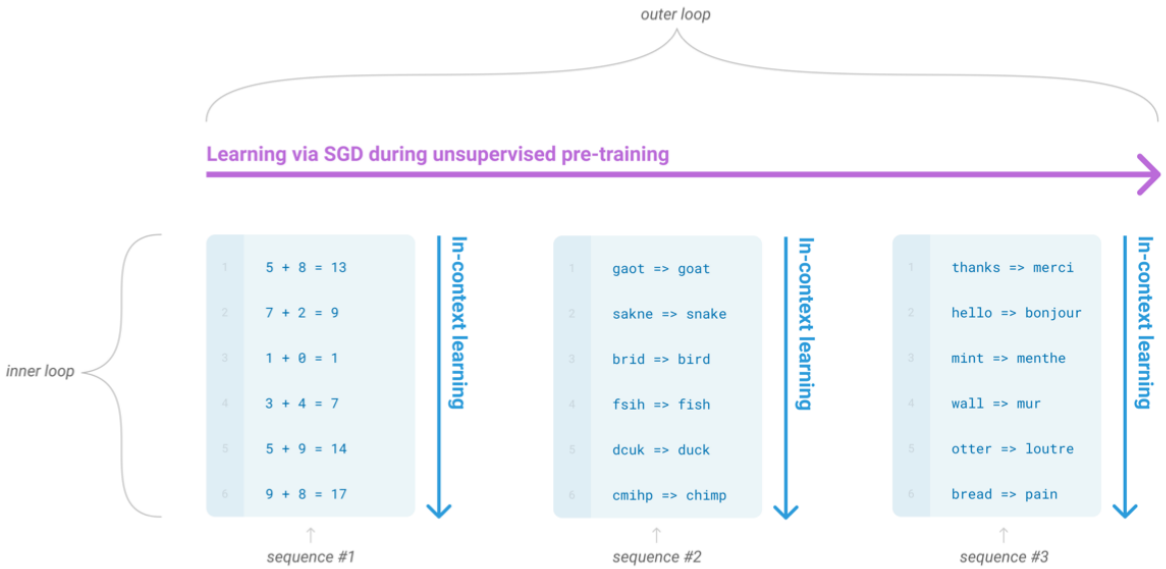


Figure 1.1: Language model meta-learning. During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training, but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence.

Few-shot learning
Chain-of-Thought

Use of Artificial Intelligence in Creativity Assessment

How does AI models have been applied in creativity assessment ?

Application pipelines

- Word embeddings
  - Unsupervised learning (BERTopic)
  - Supervised learning (fine tunning BERT)
    - Get a pre-trained BERT (general knoledge of language) run a fine-tunning in specific dataset (supervised training), that adjust the model for a specific purpose , Scoring tests, classifying texts (content analysis)

- GPT’s few-shot learning
  - Prompt engineering + few examples

Why embeddings are important for creativity assessment?

Creativity tasks ask for **novel ideas** (e.g., “What are unusual uses for a brick?”)

Traditionally, **human judges** score responses for originality

LLMs + Embeddings Can Help Automate This

Transformers like **BERT** create **context-aware word embeddings**

We can compute **Semantic Sistance** (SEM) between the prompt ("brick") and each response

Example: *"a one-use TV remote shut-off button"* → high distance = **high originality**

Deeper Measures of Divergence: **Divergent Semantic Integration (DSI):**

Measures how **distant concepts** are integrated within a response

Lower average semantic similarity = **more divergent ideas combined**

Semantic Clustering: **Eigen Entropy**

Count how many distinct **semantic clusters** appear in the response

More clusters = greater **ideational flexibility**

Validation:
These embedding-based metrics show strong correlation (~0.70) with **human originality ratings (**Beaty et. al. , 2021, Johnson et. al. 2023)

Few-shot prompting / in context learning

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam
Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss
Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh
Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

Abstract

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

1 Introduction

NLP has shifted from learning task-specific representations and designing task-specific architectures to using task-agnostic pre-training and task-agnostic architectures. This shift has led to substantial progress on many challenging NLP tasks such as reading comprehension, question answering, textual entailment, among others. Even though the architecture and initial representations are now task-agnostic, a final task-specific step remains: fine-tuning on a large dataset of examples to adapt a task agnostic model to perform a desired task.

Recent work [RWC+19] suggested this final step may not be necessary. [RWC+19] demonstrated that a single pretrained language model can be zero-shot transferred to perform standard NLP tasks

*Equal contribution
†Johns Hopkins University, OpenAI

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Meta-learning or in-context learning

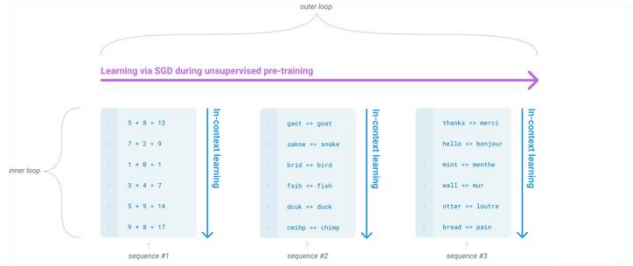


Figure 1.1: Language model meta-learning. During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training, but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence.

Few-shot learning
Chain-of-Thought

LLMs can be **prompted with rubrics + examples** to simulate expert judgment.

With few-shot or chain-of-thought prompts, they can **explain why** a response is original or appropriate

LLMs can simulate human judgment in creative scoring—not by being creative in themselves, but by leveraging vast semantic knowledge and probabilistic cues to recognize, classify, and explain creative responses.

- ✅ Validation studies show **.70–.80 correlation** with human ratings (Organisciak , et. al. 2023)

Prompt example

You are an expert teacher specialized in assessing creative writing, particularly skilled at evaluating originality and creativity in stories.

Task
Students were asked to write an original and creative story idea for a book with this cover image:

"A fantastic landscape with a sky that blends from blue to orange, under which a large full moon hovers. In the foreground, an astronaut stands next to a silver spaceship with yellow details. The ship rests on rocky terrain. In the background, imposing mountains rise, with silhouettes of trees projecting against the sky. Behind the mountains, there appears to be the silhouette of a castle or similar structure."

Your job is to evaluate and score these story ideas based on appropriateness, relevance to the image, and originality.

Scoring Rubric

Score 0: Inappropriate
- Not on task: Does not contain a coherent story idea that describes a possible plot for the book cover image.
- OR not on topic: Story has no clear connection to the key elements in the image (astronaut, spaceship, landscape features, etc.).

Score 1: Appropriate but Conventional
- On task: Outlines a coherent story idea that describes a possible plot.
- On topic: Makes explicit or implicit connections to image details (astronaut, spaceship, sky, landscape, etc.).
- Content uses familiar tropes or common themes in science fiction/fantasy (e.g., standard exploration narrative, typical alien encounter, predictable time travel story).

Score 2: Appropriate and Original
- Meets all criteria for Score 1.
- Demonstrates creative thinking with unconventional themes, unexpected plot elements, or unique perspectives.
- Shows originality that distinguishes it from typical responses (e.g., subverts genre expectations, combines unexpected elements, offers fresh interpretations of the visual elements).
- Provides thoughtful elaboration of the creative concept.

Examples
{examples_per_language["eng"]}

Evaluation Process
1. Read the story idea completely.
2. Check for connections to specific image elements (astronaut, spaceship, landscape features).
3. Assess whether the story uses conventional tropes or demonstrates original thinking.
4. Ignore grammar, spelling, or practicality issues - focus solely on creativity and relevance.
5. Assign the appropriate score based on the rubric.

Output Format
Provide scores in JSON format:

```json

{{
  "responses": [
  {{
    "id": 1,
    "score": 1
  }},
  {{
    "id": 2,
    "score": 0
  }},
  {{
    "id": 3,
    "score": 2
  }}
  ]
}}

```
Responses to Score:
{formatted_responses}
•

Does these methods work scoring using LLM works for Creativity Assessment Tasks?

Goal

Test test the convergence of AI with human raters of two methods: embeddings and few-shot learning

Method

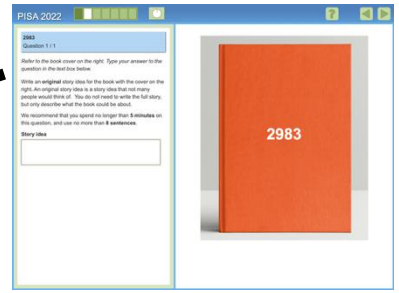
We focused on the unit titled "Book Covers," which falls under the facet "Generate Creative Ideas."

- A subset of 13,392 responses in three languages English, German and Portuguese
- Scored by trained raters (two or more) as part of the PISA

Semantic Measures from Embeddings

Semantic Distance Measures (sem_dis) :

Stimulus verbal description



Student Responses

2,983 people who caught a virus and turned into zombies. As the book unfolds, the story will go through all the details of how it happened.

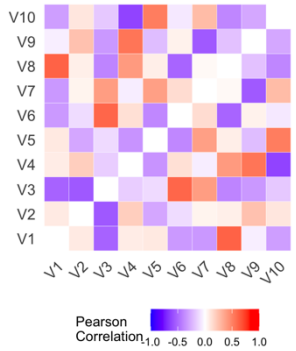
A plain, solid-orange hardcover book standing upright. In the middle of its front, the number “2983” is printed in simple white letters. There are no pictures or other words—just that bright orange surface and the white number, giving it a clean, modern look.

[0.1, 0.43,.... -0.15, 0.29, -0.12] ⟷ [0.93, 0.09,.... 0.4, -0.57, 0.95]

sem_dis = .56

Semantic distance = how far the wo texts are semantically distant (1 = distant to 0 colse)

Matrix of semantic distance of each word with all other words from the student response



Divergent Semantic Integration (dsi): the mean of these numbers sem_dis of the cell of the matrix, higher values meaning more semantically distant words within the story.

Eigen Entropy (λe): index the number of groups of related words (topics) within the idea, lower values indicate one big topic, high values several topics

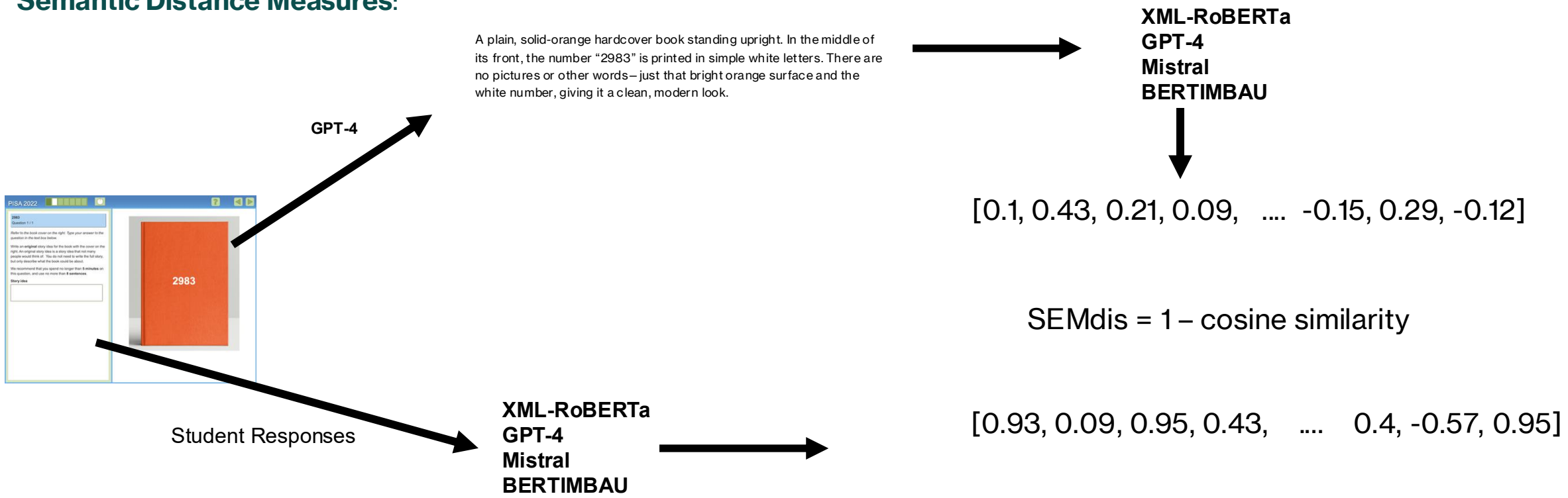
Does these methods work scoring using LLM works for Creativity Assessment Tasks?

Method

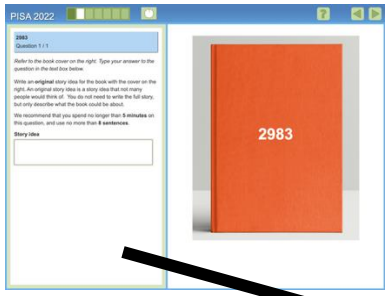
We focused on the unit titled "Book Covers," which falls under the facet "Generate Creative Ideas."

- A subset of 13,392 responses in three languages English, German and Portuguese
- Scored by trained raters (two or more) as part of the PISA

Semantic Distance Measures:



Divergent Semantic Integration (DSI)

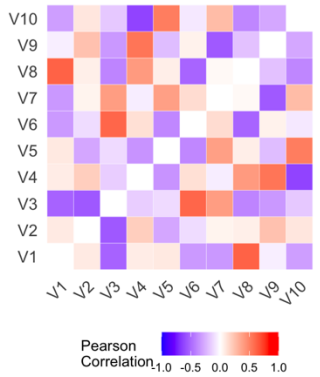


Student Responses

Token level embeddings

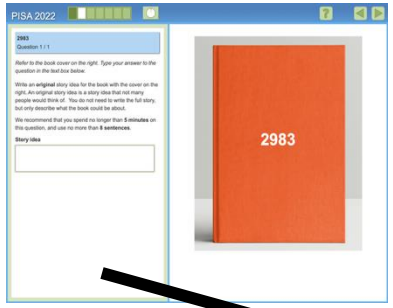
XML-RoBERTa
GPT-4
Mistral
BERTIMBAU

Matrix of cosine similarity Between tokens



DSI = average of 1-cosim of the lower triangle matrix

Refined index of semantic divergence: Eigen Entropy (λe)

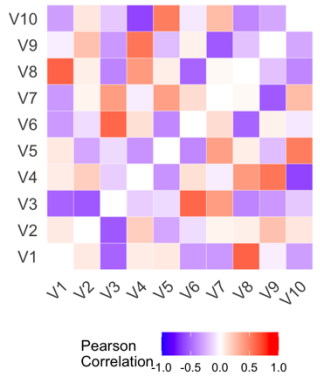


Student Responses

Token level embeddings

XML-RoBERTa
GPT-4
Mistral
BERTIMBAU

Matrix of cosine similarity Between tokens



First *n* Eigenvalues

***n* = 5**

Convert to probabilities

Each P(x) is the proportion of the mass covariance (shared semantic meaning) condensed into one component

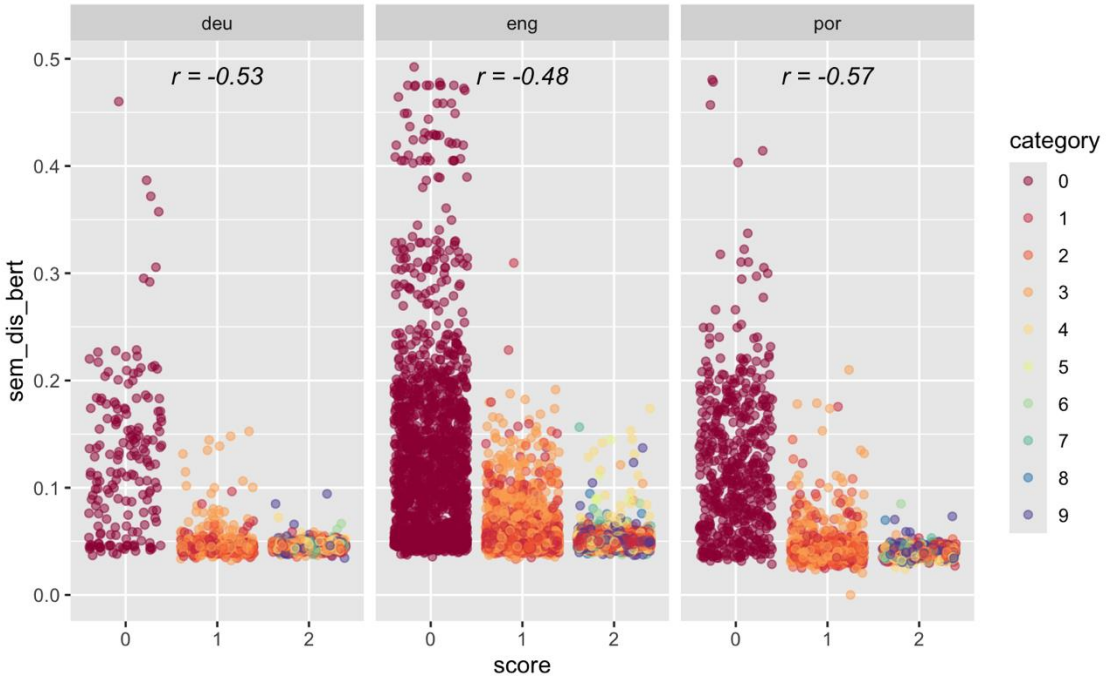
Entropy

$$\lambda e(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

High -> uniform distribution -> unpredictability -> diverse semantic components in the response

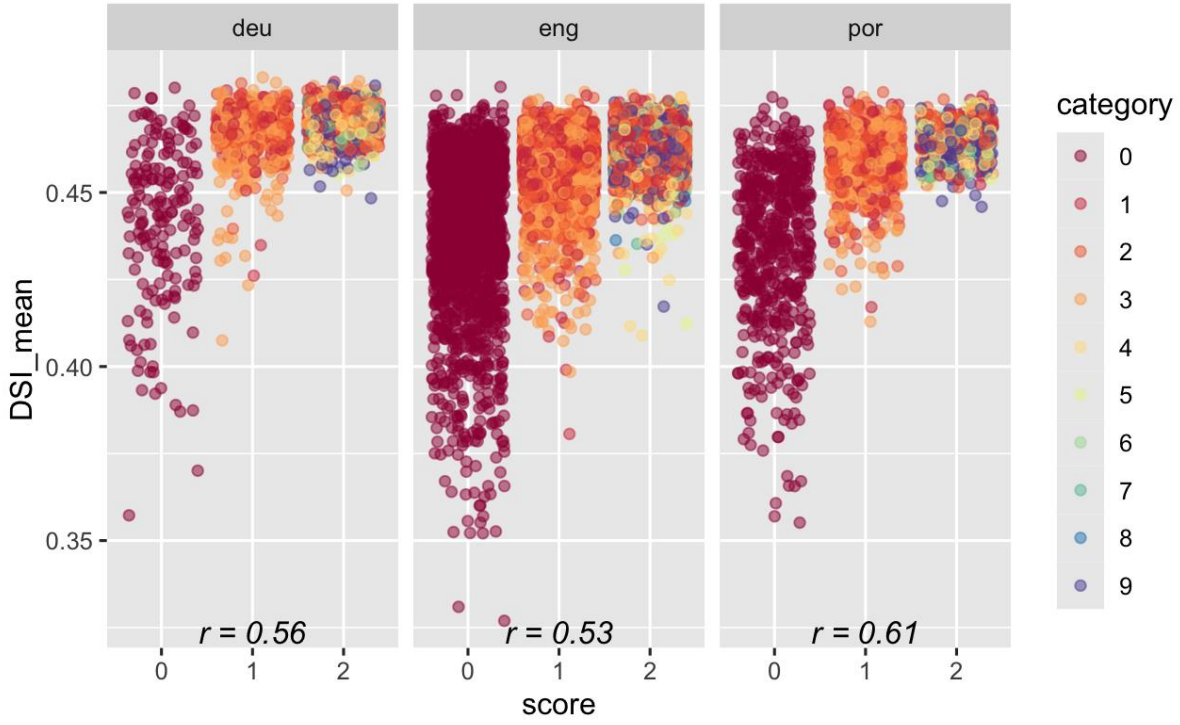
Results

Semantic Distance



Inappropriate responses are more semantically distant

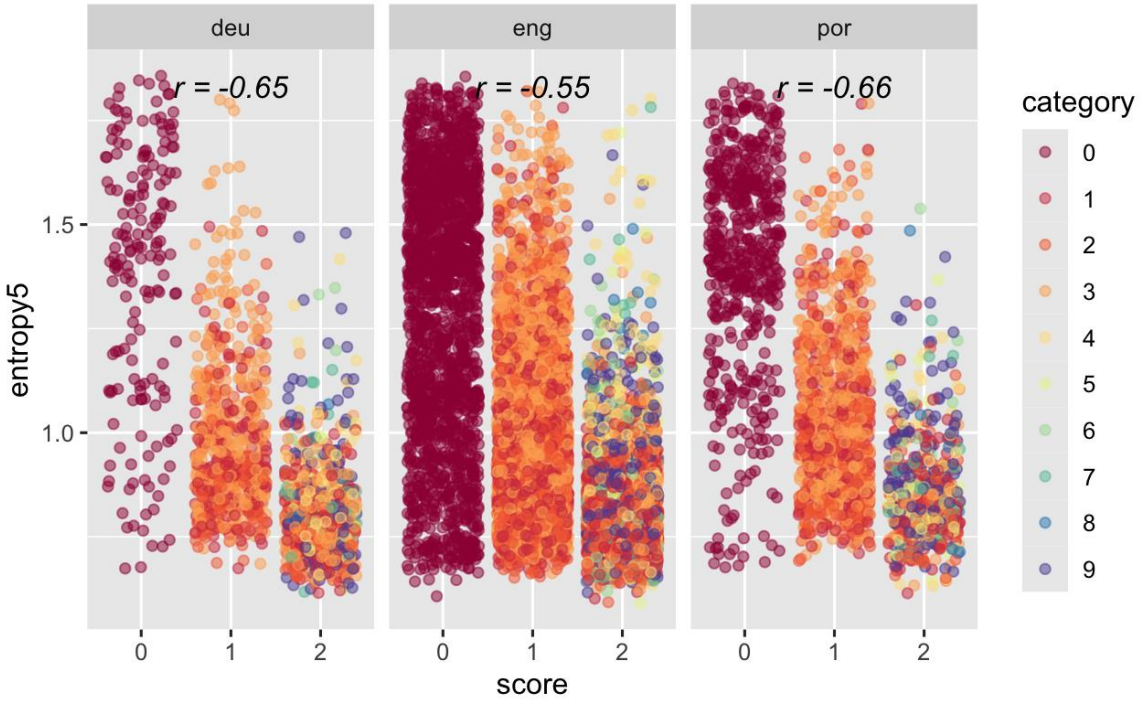
Divergent semantic Integration



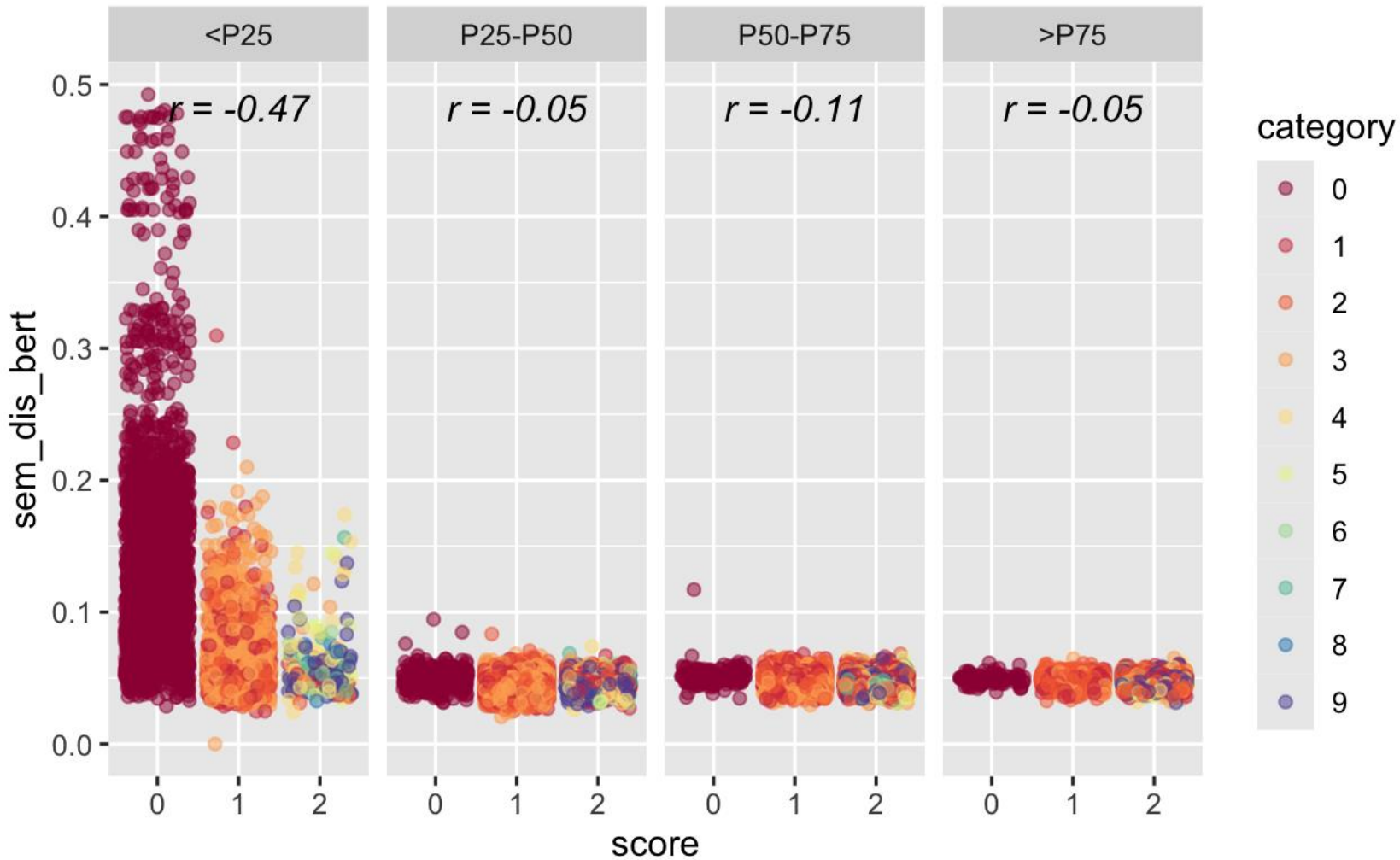
Original ideas show high divergence

Results

Eigen Entropy (λe)

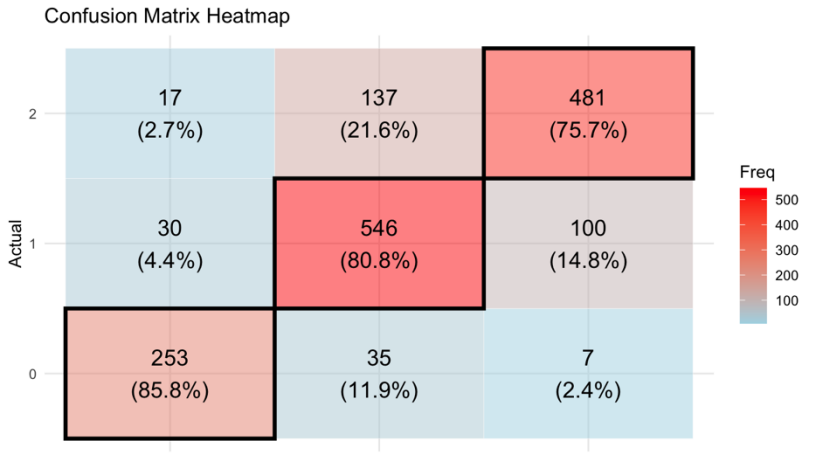


Original ideas less entropy, words cluster around one big topic (cohesion ?)


<P25
P25-P50
P50-P75
>P75
r = -0.47
r = -0.05
r = -0.11
r = -0.05
sem_dis_bert
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
score
category
0
1
2
3
4
5
6
7
8
9

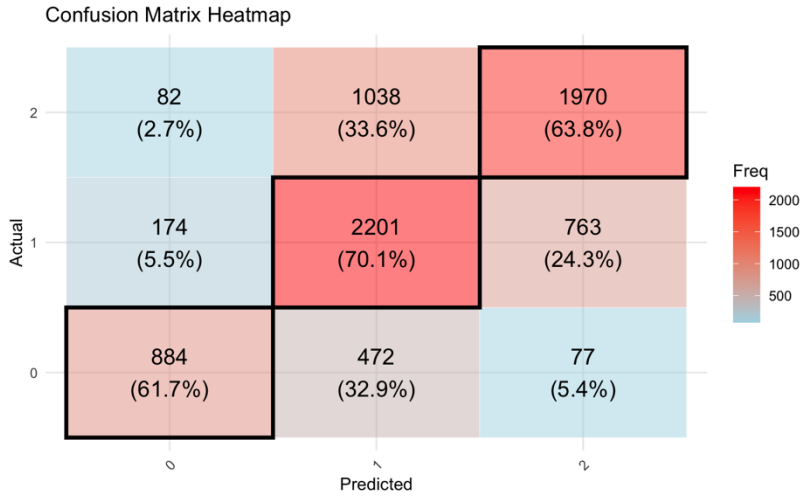
Human X Human

Metric	Value
Agreement	0.7970
Cohen's quadratic Kappa	0.7653
Macro-average Precision	0.8073
Macro-average Recall	0.8076
Macro-average F1	0.8068
Person r	0.7700



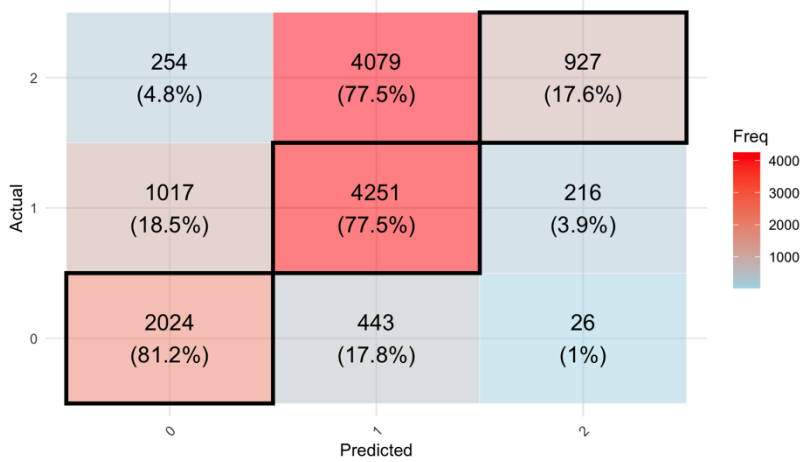
Human X gpt 4.1

Metric	Value
Agreement	0.6598
Cohen's quadratic Kappa	0.6022
Macro-average Precision	0.6899
Macro-average Recall	0.6519
Macro-average F1	0.6659
Person r	0.6000



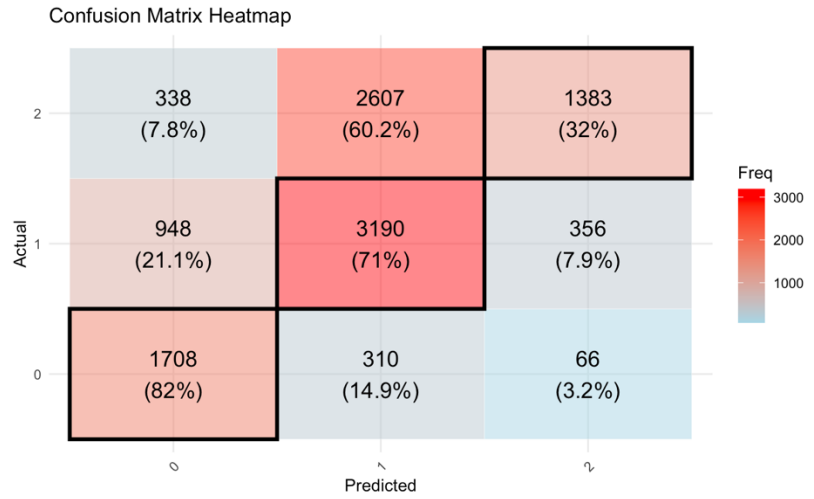
Human X Deep Seek V3

Metric	Value
Agreement	0.5441
Cohen's quadratic Kappa	0.4754
Macro-average Precision	0.6306
Macro-average Recall	0.5878
Macro-average F1	0.5280
Person r	0.5700



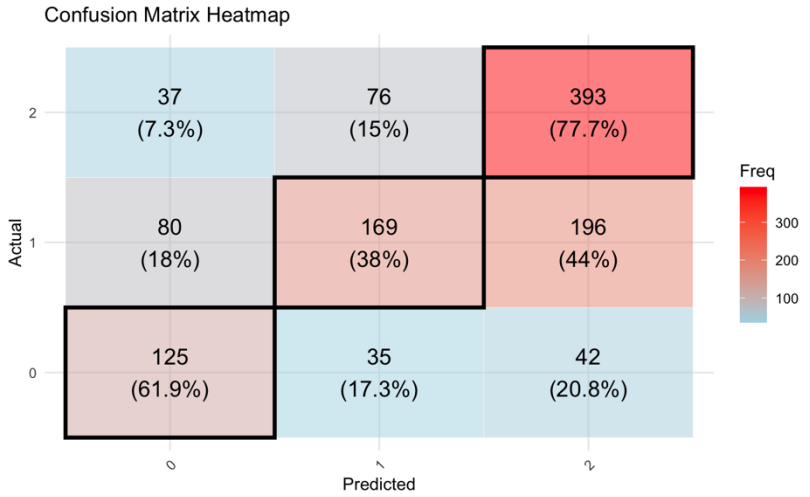
Human X gpt o3-mini

Metric	Value
Agreement	0.5759
Cohen's quadratic Kappa	0.5011
Macro-average Precision	0.6197
Macro-average Recall	0.6163
Macro-average F1	0.5752
Person r	0.5600



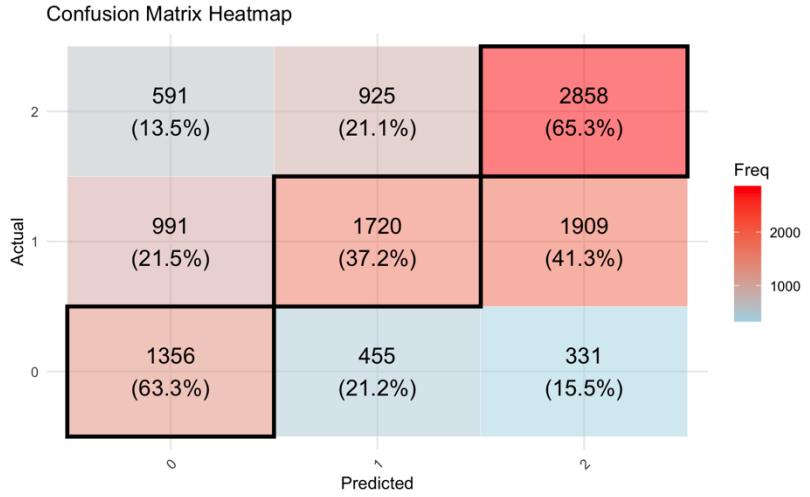
Human x GPT 4

Metric	Value
Agreement	0.5958
Cohen's quadratic Kappa	0.4891
Macro-average Precision	0.5810
Macro-average Recall	0.5918
Macro-average F1	0.5735
Person r	0.4900



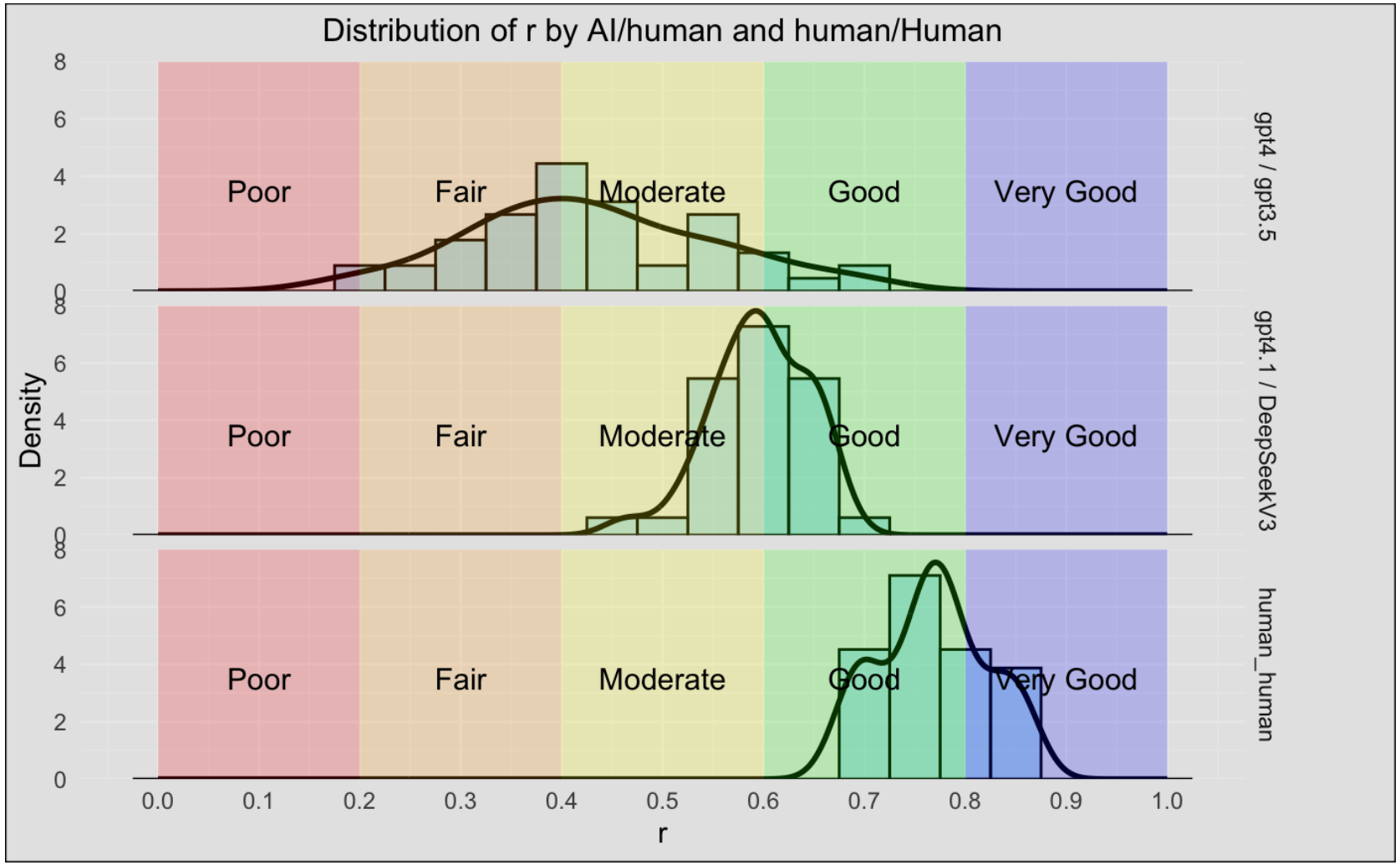
Human x GPT 3.5

Metric	Value
Agreement	0.5329
Cohen's quadratic Kappa	0.4178
Macro-average Precision	0.5257
Macro-average Recall	0.5529
Macro-average F1	0.5276
Person r	0.4200



Study results

AI-generated combined measure vs. typical rater reliability



2023

2025

Study Results

	score					
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*p*	*std. p*
(Intercept)	3.63	-0.09	2.79 – 4.47	-0.19 – 0.00	**<0.001**	0.053
sem dis bert0	-2.24	-0.35	-2.90 – -1.57	-0.58 – -0.11	**<0.001**	**0.004**
word count	0.00	0.00	0.00 – 0.00	-0.10 – 0.10	**<0.001**	0.974
DSI mean	-4.25	-0.09	-5.80 – -2.69	-0.12 – -0.06	**<0.001**	**<0.001**
score pred 3 5	0.08	0.10	0.07 – 0.10	0.08 – 0.11	**<0.001**	**<0.001**
entropy5	-0.58	-0.21	-0.67 – -0.49	-0.24 – -0.18	**<0.001**	**<0.001**
score deep seek v3	0.14	0.10	0.11 – 0.17	0.08 – 0.13	**<0.001**	**<0.001**
score gpt4 1	0.20	0.18	0.17 – 0.23	0.16 – 0.21	**<0.001**	**<0.001**
sem dis bert0 × word count	-0.09	-0.16	-0.18 – -0.00	-0.33 – -0.00	**0.049**	**0.049**
Observations	6335					
R^2 / R^2 adjusted	0.524 / 0.524					

***r* = 0.72**

Error analysis

GPT 4.1: **0**

Human: **2**

“there has been a zombie apocalypse. only one scientist can find he cure. he is sent to area 51 by the goverment to create the cure to restore human kind.”

“2042 jet racing racing in 2042 is now with jets not cars”

“Ugly emotions, beautiful pictures: In a world where objects can talk, laugh, and enjoy themselves, experience human emotions, there is only one thing that envies the life on earth. The moon, who always wanted to be able to enjoy itself with others. However, the moon is lonely and envies other rocks on the planet. One astronaut, looks at the perculiar science behind tidal waves, and the patterns it looks at, before realising that these affects that we believed to be scientific, is nothing but times when the moon gets extremely jealous. The sky, painted in a hue of beautiful colours, represents the emotions of the moon. Yellow to orange, anger. Pink to purple, the acceptance. Blue and black, the sadness. Who knew such beauty could come from sadness and ugly emotions.”

"This story it should be about astronaut who nwent on mars with his jet,he wanted to see the moon”

Error analysis

GPT 4.1: **2**

Human: **0**

“The first human to go through a black hole and end up in another dimension and end up in a planet called jpg-5 and his mission is to go back home the time fells diffrent here.”

“With Earth on the brink of collapse a brave astronomers only hope to find a solution to save mankind is to traverse through a black hole. What he finds out could change everything”

“"A man who wakes up trapped in his childhood dream — how can he escape this place? How did he end up there? What is this place? These are questions we'll explore throughout the book.""

Cost

For scoring 13,392 responses:

1,273M input tokens

1,125M output tokens

Text tokens — Price per 1M tokens · Batch API price

Model	Input	Cached input	Output
gpt-4.1 ↳ gpt-4.1-2025-04-14	$2.00	$0.50	$8.00
gpt-4.1-mini ↳ gpt-4.1-mini-2025-04-14	$0.40	$0.10	$1.60
gpt-4.1-nano ↳ gpt-4.1-nano-2025-04-14	$0.10	$0.025	$0.40

Pricing Details

MODEL[1]		deepseek-chat	deepseek-reasoner
CONTEXT LENGTH		64K	64K
MAX COT TOKENS[2]		-	32K
MAX OUTPUT TOKENS[3]		8K	8K
STANDARD PRICE (UTC 00:30-16:30)	1M TOKENS INPUT (CACHE HIT)[4]	$0.07	$0.14
	1M TOKENS INPUT (CACHE MISS)	$0.27	$0.55
	1M TOKENS OUTPUT[5]	$1.10	$2.19

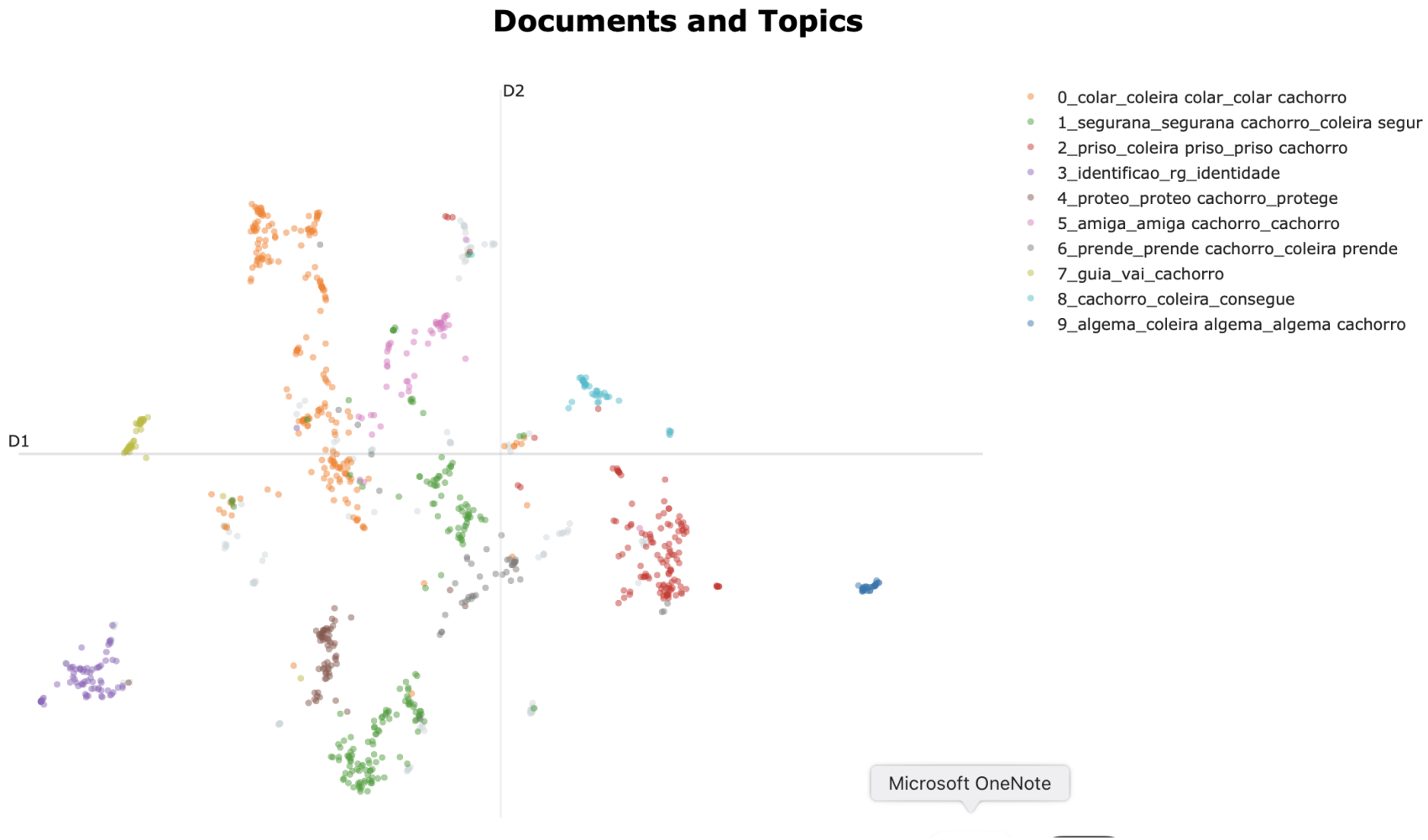
Scoring Creativity at Scale: AI Evaluation of Student-Generated Metaphors

- Study with Metaphor Creation Test (MCT)
- LLM-generated scores
  - converge with human ratings?
  - replicate patterns of external validity ?
  - show internal consistency across repeated scoring trials (i.e., model reliability under identical conditions).

Method

- **Fine-tuned model** – A Portuguese BERT model (BERTimbau; Souza, Nogueira & Lotufo, 2020) was fine-tuned to predict scores from 0 to 3.
- **Few-shot prompting** – GPT-family models (OpenAI gpt-4o, gpt-4o-mini, gpt-o3-mini, gpt-4.1), and DeepSeek models (V3 and R1) were tested using 3 to 5 prompt-response examples per item.
- **Samples**
- *Model training sample*: 12,174 metaphor responses from 974 middle and high school students, scored by 18 trained raters. Responses were scaled using the Many-Facet Rasch Model (MFRM, Primi, 2014; Primi et al 2019) to account for rater severity/leniency. 82% (9,983 responses) were used for training, and 18% (2,191 responses) for validation.
- *External validity sample*: 13,088 12th-grade students from four Brazilian states, who completed the MCT (5 items), standardized Portuguese and Math assessments, a logical reasoning test, SENNA's (Primi, et. al 2021) Openness to Experience scale, and RIASEC interest measures (Ambiel, et. al 2018).

Bertopic for content analysis



Results

Model	Agreem.	Kappa	Macro-Average F1	*Pearson r* (idea)	*Pearson r* (subject)
finetuned model					
BERT (BERTimbau)	73%	.37	.47	.47	.72
few-shot					
gpt-4o	76%	.55	.59	.57	.86
gpt-4o-mini	76%	**.58**	**.61**	**.58**	.89
gpt-4.1	73%	.48	.60	.49	.66
gpt-o3-mini	72%	.48	.52	.48	.82
gpt-o4-mini	**77%**	.54	.60	.56	.88
DeepSeek V3	76%	.57	.48	.57	**.92**
DeepSeek R1	66%	.34	.40	.35	.75

- Additionally, model stability was assessed via a test-retest protocol.
- Using the same input and temperature set to 0, the gpt-4o-mini model showed only moderate consistency (r = .77)!!
- However, rerunning the test with stricter generation parameters (fixed seed, temperature = 0, top-k = 1, and single-response batching) improved reliability to r = .90.

Validity

- Openness: Intellectual Curiosity (*r* = .19), Creative Imagination (*r* = .17), Aesthetics (*r* = .18), Tolerance to Diversity (*r* = .24)
- Cognitive outcomes: Logical Reasoning (*r* = .41), Portuguese (*r* = .33), Math (*r* = .26)
- Interest Measures: Realistic Interest (*r* = -.10) and Investigative interest (*r* = .10).

Discussion an conclusions

Initial findings

- **Reliability of AI Scoring**
  - High correlation between AI scores and human judgments.
  - Consistent performance across various language groups.
  - Presents a dependable alternative for augmenting humanscoring process
- **Advantages of Diverse Methodologies**
  - Integration of semantic distance and few-shot learning approaches.
  - Newer models GPT-4.1 / DeepSeekV3 are approaching industry standards
- **Next steps**
  - Investigating the impact of fine-tuning and prompt engineering.
  - Other methods (break the rubric and use agents to score each criteria)
  - Broadening the analysis framework to encompass all verbal creative thinking items in PISA.
  - Assessing originality in visual creative thinking tasks

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

Benedek, M. (2024). On the relationship between creative potential and creative achievement: Challenges and future directions. *Learning and Individual Differences*, 110, 102424. https://doi.org/10.1016/j.lindif.2024.102424

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The Road to Creative Achievement: A Latent Variable Model of Ability and Personality Predictors. *European Journal of Personality*, *28*(1), 95–105. https://doi.org/10.1002/per.1941

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., van Hell, J., Kennedy, E., Sullivan, G. F., Taylor, C. L., Ward, T., & Beaty, R. E. (2023). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. Behavior Research Methods, 55(7), 3726–3759. https://doi.org/10.3758/s13428-022-01986-2

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond Big and Little: The Four C Model of Creativity. Review of General Psychology.

Linacre, J. M. (1989). Many-Facet Rasch Measurement. Chicago: MESA press.

Moon, K., Green, A., & Kushlev, K. (2025, March 10). Homogenizing Effect of a Large Language Model (LLM) on Creative Diversity: An Empirical Comparison of Human and ChatGPT Writing. https://doi.org/10.31234/osf.io/8p9wu_v2

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity, 49, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (223). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity, 49, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Plucker, J. A., Beghetto, R. A., & Dow, G. (2004). Why isn't creativity more important to educational psychologists? Potential, pitfalls, and future directions in creativity research. *Educational Psychologist*, 39, 83-96.

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. The Psychology of Aesthetics, Creativity, and the Arts, 13(2), 144-152. https://doi.org/10.1037/aca0000227

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2, 68-85.

Sternberg, R. J., & O'Hara, L. A. (1998). Creativity and Intelligence. In R. J. Sternberg (Ed.), *Handbook of Creativity* (pp. 251–272). Cambridge University Press. https://doi.org/10.1017/CBO9780511807916.015

Sternberg, R. J., & Lubart, T. I. (1991). *An investment theory of creativity and its development*. Human Development, 34(1), 1–31.

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. Psychological Bulletin, 50, 362–370.

Tourangeu, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. Cognition, 11, 203-204.