

**MULTIMODAL ANALYSIS OF MEMBER INTERACTIONS
ON THE TRY GUYS CHANNEL**

THESIS

By :

Fila Izul Miya

NIM 210302110046



**DEPARTMENT OF ENGLISH LITERATURE
FACULTY OF HUMANITIES
UNIVERSITAS ISLAM NEGERI MAULANA MALIK
IBRAHIM MALANG**

2025

**MULTIMODAL ANALYSIS OF MEMBER INTERACTIONS
ON THE TRY GUYS CHANNEL**

THESIS

Presented to :

Universitas Islam Negeri Maulana Malik Ibrahim Malang

in Partial Fulfillment of the Requirements for the Degree of Sarjana Sastra (S.S.)

By :

Fila Izul Miya

NIM 210302110046

Advisor :

Drs. H. Djoko Susanto, M.Ed., Ph.D.

NIP: 196705292000031001



DEPARTMENT OF ENGLISH LITERATURE

FACULTY OF HUMANITIES

UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG

2025

STATEMENT OF AUTHORSHIP

I state that the thesis entitled **“Multimodal Analysis of Member Interactions on The Try Guys Channel”** is my original work. I do not include my materials previously written by another person except those cited as references and written in the bibliography. If there is an objection or claim, I am the only responsible for that.

Malang, 2 - Juni - 2025

The Researcher,



Fila Izul Miya
NIM. 210302110046

APPROVAL SHEET

This is to certify that Fila Izul Miya thesis entitled "**Multimodal Analysis of Member Interactions on The Try Guys Channel**" has been approved for thesis examination at the faculty of humanities, Universitas Islam Negeri (Wahid Haryani) Malik Ibrahim Malang, as one of the requirements for the degree of ~~Sarjana Sastra~~ *Sastra* (S.S).

Malang, June 23rd, 2025

Approved by :

Advisor

Head of Department of English
Literature



Drs. H. Djoko Susanto, M.Ed., Ph.D.
NIP: 196705292000031001



Ribut Wahyudi, M.Ed., Ph.D.
NIP: 198112052011011007

Acknowledge by :

Dean of Faculty of Humanities




Dr. M. Faisol, M.Ag
NIP: 197411012003121003

LEGITIMATION SHEET

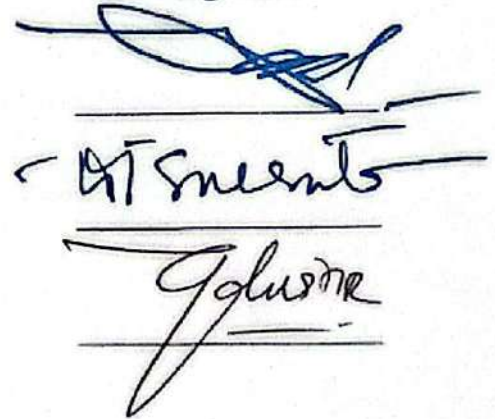
This is to certify that Fila Izul Miya thesis entitled "**Multimodal Analysis of Member Interaction on The Try Guys Channel**" has been approved for thesis examination at the faculty of humanities, Universitas Islam Negeri Maulana Malik Ibrahim Malang, as one of the requirements for the degree of Sarjana Sastra (S.S) in Department of English Literature.

Malang, 23 - Juni 2025

Board of Examiners

1. **Prof. Dr. H. Mudjia Rahardjo, M.Si.**
NIP: 195901011990031005
2. **Drs. H. Djoko Susanto, M.Ed., Ph.D.**
NIP: 196705292000031001
3. **Dr. Hj. Galuh Nur Rohmah, M.Pd. M.Ed.**
NIP: 197402111998032002

Signature



Approved by :

Dean of Faculty of Humanities




Dr. M. Faisol, M.Ag
NIP: 197411012003121003

MOTTO

لَا يُكَلِّفُ اللَّهُ نَفْسًا إِلَّا وُسْعَهَا

“Allah does not burden a person but according to his ability”

(Q.S. Al-Baqarah:286)

خَيْرُ النَّاسِ أَنْفَعُهُمْ لِلنَّاسِ

“The best of men are those who are most beneficial to others.”

(HR Ath-Thabari)

*“It wasn't just knowledge that I gained, but also lessons about life
and weathering the storm”*

DEDICATION

With heartfelt gratitude, I dedicate this thesis to my beloved parents, Mr. Sehadi and Mrs. Khoirul Uma. To my dear sister Lutfiyah, S.Ak, my cherished younger sister Dewi Husna Nabila, and my precious first nephew Muhammad Alzam Jalaluddin, thank you for being a constant source of joy and inspiration. Thank you for all your prayers, motivation, sacrifices, advice, and love that have never ceased until now.

ACKNOWLEDGEMENTS

First of all, praise Alhamdulillahirobbil'alamin. Of course, I offer my Gratitude and appreciation to the presence of Allah SWT for all His grace and gifts, which gave me ease, enthusiasm, and smoothness in completing my thesis entitled "**Multimodal Analysis of Member Interactions on The Try Guys Channel.**" Second, I send my prayers and greetings to our leader, the Prophet Muhammad SAW, who has brought us from the dark to the bright ages.

Bismillahirrahmanirrahim. There is no page more beautiful in this thesis report than the dedication sheet. With deep gratitude and heartfelt sincerity, I dedicate this thesis to the people who have played significant roles in my academic journey. First, to Prof. Dr. M. Zainuddin, M.A., the Rector of Maulana Malik Ibrahim State Islamic University Malang, and Dr. M. Faisol, M.Ag., the Dean of the Faculty of Humanities, for their leadership and the academic environment they have fostered. I also express my appreciation to Mr. Ribut Wahyudi, M.Ed., Ph.D., Head of the English Literature Department, for his dedication to the department and its students.

A special note of gratitude goes to my thesis advisor, Drs. H. Djoko Susanto, M.Ed., Ph.D., for his invaluable support, advice, and endless patience throughout this writing process. My thanks also go to Prof. Dr. Rohmani Nur Indah, M.Pd., my thesis proposal seminar lecturer, for her constructive feedback and encouragement during the early stages of this research. I am equally thankful to Mrs. Rina Sari, M.Pd., my academic advisor, for her sincere assistance and motivation throughout my studies.

From the bottom of my heart, I dedicate this thesis to the two most meritorious people in my life, Mr. Sehad and Mrs. Khoirul Uma—my beloved parents—who have never ceased praying for me and cheering me on to finish this thesis. I also dedicate this work to my beloved sisters, Lutfiyah, S.Ak, and Dewi Husna Nabila, for their unwavering love, support, and care, both emotionally and

materially. To my dear nephew, Muhammad Alzam Jalaluddin, thank you for bringing joy and warmth into our family.

To my circle of support, Escanor'21, thank you for walking through this chapter of life with me. I'm grateful for the memories and companionship we shared. To Afrilia Nur Chasanah, Tazkia Zahra Bukhori, and Farha Abidah, I am especially thankful for your constant support, cheerful presence, and the journey we've been through together. Each challenge felt lighter, and each moment brighter with you by my side.

Lastly, I want to dedicate this thesis to myself. Thank you for being a part of your own happiness—for working hard, for not giving up, and for believing that you could make it through all of this. Thank you for staying true to yourself and for persevering through pressure and doubt. This achievement is something to be truly proud of—a symbol of strength, growth, and self-love.

Malang, 2 Juni.... 2025

The Researcher,



Fila Izul Miya

NIM. 210302110046

ABSTRACT

Miya, Fila Izul (2025). *Multimodal Analysis of Member Interaction on The Try Guys Channel.* Undergraduate Thesis. Department of English Literature, Faculty of Humanities, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisor, Djoko Susanto, M.Ed., Ph.D.

Keyword: *Multimodal, The Try Guys, Youtube*

This study explores how multimodal communication is constructed through verbal, visual, gestural, audio, and spatial elements among members of The Try Guys, a popular American YouTube group known for their humorous and interactive content. Using Kress & van Leeuwen’s Multimodal Discourse Analysis as the theoretical framework, this research investigates the coordination between language and other semiotic modes in the video titled “Can We Guess This Celebrity Perfume? – Common Sense.” The findings show that humor, group identity, and emotional expression are not solely produced by verbal language but are significantly enhanced through synchronized non-verbal and visual cues, such as facial expressions, bodily gestures, intonation, spatial arrangement, and visual design. These multimodal features serve both representational and interpersonal functions, enriching meaning-making in digital communication. By analyzing the interaction patterns among the group members, this research reveals how digital content creators effectively engage their audience through a layered mode of storytelling. The study concludes that multimodal communication is not merely a supplementary element but a core component of meaning construction in online media. This thesis contributes to the academic understanding of multimodal discourse and offers practical implications for media practitioners aiming to optimize viewer engagement on digital platforms and social media channels.

المخلص

ميا، فيلا إزول (2025). تحليل متعدد الوسائط لتفاعل الأعضاء في قناة "ذا تراي غايز". رسالة البكالوريوس، قسم اللغة الإنجليزية، كلية العلوم الإنسانية، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف: د. دجوكو سوسانتو، ماجستير في التربية، دكتوراه.

الكلمات المفتاحية: التحليل متعدد الوسائط، ذا تراي غايز، يوتيوب

تناقش هذه الدراسة كيفية بناء التواصل متعدد الوسائط من خلال العناصر اللفظية، البصرية، الإيمائية، الصوتية، والمكانية في التفاعل بين أعضاء "ذا تراي غايز"، وهي مجموعة يوتيوب أمريكية شهيرة معروفة بمحتواها الفكاهي والتفاعلي. باستخدام منهجية اللسانيات النظامية الوظيفية لهاليداي، وتحليل الخطاب متعدد الوسائط لكريس وفان ليووين، تحلل هذه الدراسة التنسيق بين اللغة وأنماط السيميائية المختلفة في الفيديو المعنون: "هل يمكننا تخمين هذا تشير نتائج التحليل إلى أن الفكاهة، والهوية الجماعية، والتعبير." العطر الخاص بالمشاهير؟ - الحس المشترك العاطفي لا تُبنى فقط من خلال اللغة اللفظية، بل يتم تعزيزها بشكل كبير بواسطة العناصر غير اللفظية المتزامنة، مثل تعابير الوجه، وحركات الجسد، ونبرة الصوت، وتنظيم الفضاء، والتصميم البصري. تؤدي هذه العناصر متعددة ومن خلال دراسة أنماط التفاعل. الوسائط وظائف تمثيلية وتفاعلية، مما يُثري عملية بناء المعنى في التواصل الرقمي بين الأعضاء، تكشف الدراسة كيف ينشئ صانعو المحتوى الرقمي تفاعلاً جماهيرياً من خلال السرد متعدد الوسائط. وتخلص الدراسة إلى أن التعدد الوصائطي ليس مجرد عنصر مكمل، بل هو جانب أساسي في بناء المعنى في الوسائط الرقمية. وتُقدم هذه النتائج مساهمة في تطوير دراسات الخطاب متعدد الوسائط وتطرح دلالات عملية لصناع المحتوى الذين يسعون لتعزيز تفاعل الجمهور بشكل فعال.

ABSTRAK

Miya, Fila Izul (2025). **Analisis Multimodal terhadap Interaksi Anggota dalam Kanal The Try Guys**. Skripsi Sarjana. Jurusan Sastra Inggris, Fakultas Humaniora, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: Djoko Susanto, M.Ed., Ph.D.

Kata Kunci: *Multimodal, The Try Guys, YouTube*

Penelitian ini mengeksplorasi bagaimana komunikasi multimodal dibangun melalui unsur verbal, visual, gestural, audio, dan spasial di antara anggota The Try Guys, sebuah grup YouTube populer asal Amerika yang dikenal dengan konten mereka yang humoris dan interaktif. Menggunakan Analisis Wacana Multimodal dari Kress & van Leeuwen sebagai kerangka teoretis, penelitian ini mengkaji koordinasi antara bahasa dan mode semiotik lainnya dalam video berjudul “Can We Guess This Celebrity Perfume? – Common Sense.” Temuan menunjukkan bahwa humor, identitas kelompok, dan ekspresi emosional tidak hanya dihasilkan melalui bahasa verbal, tetapi secara signifikan diperkuat oleh isyarat nonverbal dan visual yang tersinkronisasi, seperti ekspresi wajah, gerakan tubuh, intonasi, penataan ruang, dan desain visual. Fitur-fitur multimodal ini menjalankan fungsi representasional sekaligus interpersonal, yang memperkaya proses pembentukan makna dalam komunikasi digital. Dengan menganalisis pola interaksi antar anggota grup, penelitian ini mengungkap bagaimana para kreator konten digital secara efektif melibatkan audiens mereka melalui mode penceritaan yang berlapis. Studi ini menyimpulkan bahwa komunikasi multimodal bukanlah elemen pelengkap semata, melainkan komponen inti dalam konstruksi makna di media daring. Skripsi ini memberikan kontribusi terhadap pemahaman akademik tentang wacana multimodal dan menawarkan implikasi praktis bagi pelaku media yang ingin mengoptimalkan keterlibatan penonton di platform digital dan saluran media sosial.

TABLE OF CONTENT

| | |
|---|------|
| STATEMENT OF AUTHORSHIP | i |
| APPROVAL SHEET | i |
| LEGITIMATION SHEET | iii |
| MOTTO | iv |
| DEDICATION | v |
| ACKNOWLEDGEMENTS | vi |
| ABSTRACT | viii |
| مستخلص البحث | ix |
| ABSTRAK | x |
| TABLE OF CONTENT | xi |
| CHAPTER I : INTRODUCTION | 1 |
| A. Background Of The Study | 1 |
| B. Problem Of Study | 6 |
| C. Significance Of The Study | 6 |
| D. Scope and Limitation | 7 |
| E. Definition Of Key Term | 7 |
| CHAPTER II: LITERATURE REVIEW | 9 |
| A. Multimodal | 9 |
| B. Youtube | 17 |
| CHAPTER III: METHODS | 19 |
| A. Research Paradigm | 19 |
| B. Data Source | 21 |
| C. Data Analysis | 21 |
| D. Data Collection | 22 |
| E. Research Instrument | 23 |
| CHAPTER VI: FINDINGS AND DISCUSSION | 24 |
| A. Findings | 24 |

| | |
|--|----|
| B. Discussions | 53 |
| CHAPTER V: CONCLUSION AND SUGGESTION | 62 |
| A. Conclusion | 62 |
| B. Suggestion | 64 |
| BIBLIOGRAPHY | 66 |
| CURRICULUM VITAE | |

CHAPTER I

INTRODUCTION

A. Background of tshe Study

In the current digital era, communication increasingly involves not only language, but also various other modes such as gesture, facial expressions, body movements, and visual layouts that complement each other to convey meaning. This phenomenon, known as multimodality, plays a crucial role in shaping online discourse. Multimodal is an approach that considers communication as not just language or words; it also considers a variety of other ways to convey messages, such as music, sound, images, gestures, posture, and spatial layout (Jewitt & Jones, n.d.). All these elements work together to better interact and convey messages. This approach emphasizes the importance of combining different modes of communication to create richer and deeper meanings in interactions.

Kress and van Leeuwen (1996) define multimodal as the simultaneous use of multiple modes in the design of a semiotic product or event, where these modes are integrated to reinforce, complement, or arranged in a way that creates richer meaning in communication. Meanwhile, Iedema, R. (2003) states that multimodal is a technical term that indicates that the meaning we make utilizes a variety of semiotic sources, including not only language but also images, sounds, and movements.

As a social media category, YouTube video sites are mostly used by users who are extroverted in their social lives. YouTube is one of Google's video

sharing services. It allows users to upload, watch, share and search for video clips independently and for free. YouTube is a technological shift of the internet, or world wide web, which allows users to read and share information for other users (Chakma et al., 2022). This causes YouTube to become one of the practical and easily accessible social media, so that today YouTube is a well-known site and is watched by thousands of people every day (Chakma et al., 2022) (Chen et al., 2017).

YouTube is not a “text”, but rather a website where users can view videos (Benson, 2005). YouTube provides a space for users to express themselves and interact with the audience through a variety of engaging video formats. In addition, YouTube provides a commenting system platform that allows viewers to write comments directly on the video screen and access specific scenes, thus creating a sense of “live” and shared viewing experience. The website combines visual, auditory and interactive elements, making it a dynamic medium for communication and creativity.

In the study of human communication, the multimodal approach is important because it reveals how different modes of communication work together to convey meaning. Multimodal studies how different components of communication - such as writing, spoken language, images, sounds, gestures, facial expressions and spatial arrangements - work together to provide richer and more complex meanings. Multimodality is also closely related to studies on how messages are conveyed and received in various social and cultural situations (Suparno, Thamrin, & Chairul, 2022).

To identify gaps in this study, researchers analyzed previous studies relevant to the current study. The first research by (Kadwa & Alshenqeeti, 2020) examines Indonesian and international Pantene advertisements using a multimodal approach based on Halliday's transitivity theory and Kress & van Leeuwen's social visual theory. The results show that the representation of women with strong and beautiful hair is constructed differently in the two ad versions. The findings reveal that the use of verbal and visual elements in advertisements plays a role in shaping certain cultural stereotypes, which can also be the basis for understanding how identity narratives are displayed multimodally, as happens in Try Guys members' interactions on digital platforms.

Multimodal analysis is also used in advertising contexts. For instance, (Ulati, 2021) research analyzed Energen advertisements using multimodal theory by Anstey and Bull. The ads were analyzed from linguistic, visual, audio, gestural, and spatial aspects to show how messages are communicated comprehensively. The research shows that the family interaction in the ad creates interpersonal closeness with the audience, as well as forming an understanding of the value of the product. This multimodal approach reinforces the relevance in the context of analyzing social interaction in digital media such as YouTube, where verbal and non-verbal elements are used together to convey messages and build closeness with the audience.

In addition, there are previous studies that study by (Janvier et al., 2022) conducted a multimodal analysis of the response of bioengineered tendon tissue to cyclic stress variations. The study used techniques such as qPCR, TEM, second-

harmonic imaging, and mechanical analysis to assess how different collagen isoforms play a role in the formation of tendon structures. Although the context is bioengineering, the use of a multimodal approach that combines visual, numerical and molecular data emphasizes the importance of integrating multiple forms of information to uncover complex dynamics—a principle that is relevant in studying the dynamics of multimodal interactions between members of video content such as Try Guys.

The role of multimodality in shaping digital narratives has also been examined in the context of disinformation. (Wilson et al., 2023) in their systematic literature review explored how disinformation and misinformation are constructed multimodally, especially in video and social media. The review emphasizes the importance of the interplay between text, images, sounds and gestures in influencing audience perceptions in a hidden yet powerful way. Multimodality is explained as an instrument of manipulation that amplifies the emotional effects and visual appeal of misleading information. This study provides a critical analytical framework for how messages are perceived holistically by digital audiences, very much in line with the analysis of how Try Guys members use multiple communication channels (gestures, expressions, voice, and text) in shaping their collective narrative.

Some of the previous studies, there is also (Walkington et al., 2024) introduced the Multimodal Analysis for Embodied Technologies (MAET) framework to analyze learning interactions in the context of AR, VR, and motion capture-based educational technologies. This research shows how gesture, body

position, spatial interaction, and language are used in collaborative learning, emphasizing the importance of embodied cognition in concept understanding. MAET is an ideal method for evaluating motion and visual-based social interactions-which are also important components in YouTube content such as Try Guys, where bodily and visual expressions are integral to communication between members.

The Try Guys is a popular American YouTube channel known for its creative and entertaining content. The Try Guys team incorporates elements of comedy, experimentation, and interaction in each of their videos. To convey messages in an engaging manner, they often use verbal elements, such as conversation and humor, combined with non-verbal elements, such as gestures, facial expressions, and visual editing effects (Kress & Van Leeuwen, 2001). The use of these multimodal elements allows them to convey a stronger message and strengthen the emotional bond with the audience.

Different from some previous studies, the purpose of this study aims to examine how verbal, visual, and non-verbal elements are utilized in interactions among The Try Guys members. It also explores how these elements work together to form communication patterns that reflect group dynamics and audience engagement. By applying multimodal discourse analysis, this research seeks to uncover how meaning is constructed through various modes in YouTube video content.

B. Problem of Study

Based on the background of the study, the researcher asks two questions in this study :

1. What forms of interaction do the members display in “*The Try Guys*” Content?
2. How do the members of “*The Try Guys*” use multimodal elements to build communication?

C. Significance of the Study

This study is expected to make an important contribution both theoretically and practically. Theoretically, this study expands the field of multimodal discourse analysis by using the approach developed by Kress and van Leeuwen (2006). By examining the use of verbal, visual, and non-verbal modes in interactions between members of The Try Guys channel on YouTube, this study provides new insights into how meaning is constructed through the combination of various semiotic resources in the context of digital communication.

Practically, this study is useful for digital content creators in understanding how interactions between individuals in videos can influence audience perception through the integrated use of multimodal elements. The findings of this study can also serve as a reference for academics and researchers interested in developing studies on digital communication and visual representation in social media. Thus, this study is expected to make a meaningful contribution to the development of

multimodal communication studies in the increasingly evolving era of digital media.

D. Scope and Limitation

This study employs a Multimodal Discourse Analysis to examine how language and multimodal elements—verbal, visual, and non-verbal—are used in social interactions among members of The Try Guys. The data analyzed is limited to a single video titled “*Can We Guess This Celebrity Perfume? – Common Sense*” published on *The Try Guys* YouTube channel.

The focus of this study is on the interaction between the main members of the group, excluding audience comments, production crew interactions, or feedback on other platforms. Furthermore, this research does not explore technical aspects of production such as lighting, transitions, or video editing mechanics.

The study applies Kress and van Leeuwen’s (2006) theory to understand how semiotic resources—such as gesture, posture, tone, and visual layout—interact with verbal communication to construct meaning in digital discourse. Thus, the scope is restricted to analyzing interpersonal communication within a multimodal framework, based entirely on content from the YouTube video.

E. Definition of key term

- a. Multimodality:** The integration of various modes—such as verbal language, gestures, visuals, and sound—to create meaning in communication.

- b. Member Interaction:** Communicative behavior and social dynamics among individuals within a group setting.
- c. The Try Guys:** A YouTube-based content creator group known for humorous, challenge-based, and collaborative videos.
- d. YouTube:** A video-sharing platform that enables users to create, upload, view, and interact with multimedia content across diverse genres.

CHAPTER II

LITERATURE REVIEW

A. Multimodal

Multimodal is a term used to describe how people interact by utilizing various communication channels simultaneously (Kress & van Leeuwen, 1996). Multimodal is also described as the simultaneous use of various semiotic modes in designing a product or communication event, where these modes are arranged in such a way that they support, complement, or are arranged in a specific order (Kress and van Leeuwen, 2001).

Multimodal can also be understood as a technical term indicating that humans use various sign systems or semiotic modes in the process of constructing meaning (Iedema, 2003). Multimodal analysis is an approach that involves various forms of communication by combining text and relationships between two or more semiotic elements to achieve communicative goals in a text. This approach emphasizes that all forms of communication, both verbal and nonverbal, play an important role in shaping meaning. This is because language is considered to contain meaningful information.

The changes occurring in today's society are the result of increasingly complex social transformations. In this context, meaning is formed through a combination of verbal and visual elements as modes of communication. Bezemer and Kress (2015) explain that the combination of these two modes—verbal and

visual—is at the heart of multimodality studies, where “multi” means many and “modality” refers to the means or mode of communication.

Jewitt (2009) and O’Halloran (2011) state that multimodality discourse is closely related to understanding the process of communication and representation of meaning, which involves the interaction of various semiotic resources such as language, body movements, gaze direction, and camera angles. According to Kress and van Leeuwen (1996), multimodality refers to how humans communicate using more than one mode simultaneously. Therefore, the study of multimodality emphasizes how semiotic resources are used to convey messages effectively.

Kress (2009) also explains that modes are meaningful resources shaped by social and cultural influences and used in the process of meaning formation. These modes are not limited to language but include elements such as images, visual arrangements, sound, and intonation. The combination of these various modes is often used to construct a complete message. Furthermore, Jewitt, Bezemer, and O’Halloran (2016) define mode as a term in social semiotics that refers to the resources used in the process of meaning construction. This confirms that mode is part of an organized multimodal system aimed at creating meaning.

Multimodal analysis emphasizes all forms of communication, both verbal and nonverbal, which contribute significantly to the conveyance of meaning because they contain informative value. Fang (2019) explains that the term “modality” in multimodality refers to the ways and media of communication,

which include symbolic systems such as technology, image colors, and music or sound. Language in this context does not stand alone but is part of a communication system that integrates various modes simultaneously. Multimodal communication essentially focuses on the study of various modalities used through diverse symbolic resources such as sound, symbols, images, and color.

Bull and Anstey (2018) classify types of multimodal texts into three categories.

1. Multimodal texts can be found in printed form, such as books, comics, and posters.

2. Multimodal texts are also present in digital form, including slide presentations, e-books, blogs, e-posters, websites, and social media such as Facebook, Twitter, and Instagram, which combine visual and textual elements simultaneously. This type also includes animation-based media, films, and digital games such as video games.

3. Multimodality can also appear in live performances or stage events, such as ballet and theater. These three forms show that multimodality can be present in various media and contexts, both static and dynamic.

According to the theory proposed by Anstey and Bull (2018), a text can be called a multimodal text if it contains two or more semiotic systems. A semiotic system refers to the use of signs and symbols as tools to convey meaning. Based on the explanation by Michèle Anstey and Geoff Bull in their work, there are five types of semiotic systems that form the basis of multimodal texts namely :

1. Linguistics

The linguistic system includes various elements such as vocabulary, general text structure, word choice, grammar, and sentence and paragraph structure.

a. Vocabulary

On YouTube, vocabulary appears through the words spoken by creators and supporting text (subtitles, overlays). “Gadget review” content, for example, uses technical terms such as refresh rate, chipset, or ultrawide. Meanwhile, lifestyle channels often use casual phrases like “healing tipis-tipis” or “quality time bareng bestie” to build a connection with viewers. The right choice of words helps algorithms match videos with search keywords and makes it easier for viewers to quickly understand the content.

b. Generic Structure

Popular YouTube videos almost always follow this pattern: hook \leq 15 seconds \rightarrow brief intro \rightarrow main content \rightarrow CTA (call-to-action). The hook serves the same purpose as the “promotional goal” in ads—to grab attention as quickly as possible (“See how I transformed my dorm room into a café-style space in just 3 hours!”). Product/brand mentions can be placed in the intro (“This video is sponsored by...”), in text overlays, or in the description with a link. The “product name/activity” and brand mention are shown again before the CTA: “Click the link below for a 20% discount.” This structured format helps viewers navigate the video’s content without feeling lost while also fulfilling commercial objectives.

c. Word Choice

Due to high competition, titles and thumbnails must be persuasive, concise, and logical. Example: Title: “7 Excel Tricks That Make Work 2× Faster” (persuasive & clear benefits). Description: polite language, yet still to-the-point, plus time stamps to help viewers jump to important sections. In scripts, creators often use an “inviting” language pattern (“let’s try,” “don’t forget”) to make viewers feel involved. This style aligns with the persuasive function of ads in Anstey & Bull’s theory.

d. Grammar and Organization

Indonesian YouTube creators often switch between formal and informal language—for example, “Guys, this feature literally helps a lot when working remotely.” Jargon is present in certain niches: finance channels use “dollar-cost averaging,” while gaming channels mention “nerf” or “buff.” Slang and Jakarta slang (“jujurly,” “sabi,” “BTW”) create a casual atmosphere and feel relatable to younger audiences. While not always standard, grammar must remain consistent; chaotic automatic subtitles can reduce viewer retention and accessibility. Therefore, many creators review subtitles manually to ensure the message remains credible.

Thus, all linguistic components—from word choice to grammar—serve as layers of meaning that integrate with visual elements (thumbnails, text overlays), audio (intonation, background music), gestural elements (host expressions), and spatial elements (screen layout, product placement). This multimodal combination

is what makes YouTube videos effective at captivating, informing, and influencing viewer actions.

Linguistic modes are divided into two types, namely written language and spoken language.

- a) Written language includes various forms of communication that are expressed in written form. Examples include invitation cards, newspapers and magazines, personal letters, game instructions, food packaging, and academic essays.
- b) Spoken language includes forms of communication that are conveyed directly through speech. Examples include everyday conversations, singing songs, reading books, or telling a story. In this form, a person conveys a message by speaking directly using spoken words.

2. Visual

Visual mode includes various elements that can be seen directly by the sense of sight, such as images, videos, colors, visual layout, design, font types, text size, writing format, graphics, and visual aids such as tables, diagrams, charts, concept maps, animations (GIFs), and vectors. In this mode, meaning is conveyed through the use of images and other visual symbols. Examples of visual mode include television shows or movies, still or moving images, photos, illustrations, icons, and facial expressions. These elements can convey certain emotions, such as happiness, sadness, anger, surprise, fear, or hatred. It should be noted that the

meaning of a symbol or visual sign can vary depending on the culture, which can sometimes lead to misperceptions. In addition, visual signs often appear spontaneously without prior planning. Eye contact is also an important part of nonverbal visual communication, as it can show attention or interest in the other person.

3. Audio

Audio mode includes elements such as sound volume, pitch, and rhythm in music or sound effects. In spoken communication, sound is the main element—whether it is loud, soft, or even a whisper. The sense of hearing plays an important role in capturing someone's attention, and through sound, messages can be conveyed in a more vivid and meaningful way. This mode communicates meaning through various types of sounds and sound devices in a specific communication situation. For example, in online learning videos, a clear and expressive narrator's voice helps students understand the material more easily. The use of sound effects such as a “ding” sound when important information is presented also adds appeal and emphasis. In this case, the narrator can be considered part of the multimodal text because they combine words (linguistic mode) with sound (audio mode) to convey the message more effectively.

4. Gestural

Gestural mode relates to communication conveyed through body movements, facial expressions, speed or slowness of movement, and meaningful silence. In this mode, meaning is expressed through body language as a means of conveying

emotions, attitudes, or reactions without words. For example, in a YouTube video presentation, a speaker who raises both eyebrows while explaining something may indicate amazement or surprise. Open hand gestures while speaking indicate openness and confidence, while crossing your arms over your chest can indicate defensiveness or discomfort. Expressions such as smiling when greeting the audience or nodding when listening to questions are also part of gestures that reinforce the message being conveyed verbally.

5. Spatial

Spatial mode relates to aspects such as distance (proximity), direction, position, as well as the layout and organization of objects in a space. In this mode, meaning is conveyed through the way objects or people are placed and arranged in an area, as well as how that positioning affects the audience's interpretation of the message. For example, in a cooking vlog, ingredients arranged neatly from left to right in order of use can help viewers understand the process more easily. Similarly, in digital educational content, the speaker's position in the center of the screen with explanatory text on the right side helps viewers focus visually. This strategic placement of visual elements is very important in creating efficient and easy-to-understand communication. Based on the explanation of the five semiotic systems in multimodal theory, this study chose to use this theory because it is considered capable of helping researchers understand the issues being studied and achieve the research objectives. In addition, this theory is purely multimodal, not a derivative or development of another theory, so it can provide a more direct and comprehensive approach to multimodal phenomena in communication.

B. YouTube

YouTube is more than a platform for video distribution; it is a complex digital ecosystem that enables and shapes contemporary forms of public discourse, identity performance, and multimodal communication. According to Burgess and Green (2018), YouTube operates as a “cultural archive and a participatory platform” where everyday users and professional creators alike construct meaning through media production, interaction, and audience engagement.

From a communication theory perspective, YouTube is understood as a hybrid media space a platform that blends elements of interpersonal conversation, mass broadcasting, and interactive feedback (Jenkins, Ford, & Green, 2013). It enables not only content delivery but also dynamic meaning-making, as creators can encode messages using verbal language, visual imagery, music, editing techniques, and platform-specific conventions such as thumbnails, hashtags, and end screens.

Lange (2007) highlights that YouTube fosters the emergence of “technical identities,” where creators become known not only by their personalities but also by their stylistic and aesthetic choices—such as jump cuts, typography, framing, and tone of delivery. These elements, while often non-verbal, are integral to how messages are interpreted and how social relationships are formed in digital communities.

Moreover, YouTube as a discourse space is inherently multimodal. Its architecture invites content that is inherently layered—spoken dialogue is

accompanied by visuals, music, gestural performance, and editorial sequencing. Kress and van Leeuwen (2001) emphasize that in such settings, meaning is the result of mode interaction, not of linguistic content alone.

In the context of this study, YouTube provides the semiotic stage for analyzing the interactive communication of The Try Guys. Their video content, particularly in the episode “Can We Guess This Celebrity Perfume?”, involves multimodal strategies such as eye contact with the camera (visual), exaggerated gestures (non-verbal), layered audio cues (auditory), and humorous dialogue (verbal), all of which interact to produce a cohesive communicative experience. These features exemplify how YouTube content embodies the principles of performative discourse, where creators engage not only with co-speakers but also with an imagined audience.

Therefore, understanding YouTube as a unique discursive environment is crucial in analyzing the meaning-making practices in The Try Guys’ content. The platform’s affordances, such as editing tools, comment sections, subscriber metrics, and algorithmic exposure, actively shape how creators communicate and how audiences perceive authenticity, humor, and group dynamics.

CHAPTER III

METHOD

This chapter attempts to present the research methodology used by researchers to examine and analyze research data. It is divided into several parts such as paradigm, data sources, data analysis, data collection and instruments.

A. Research Paradigm

According to Rahardjo (2018) defines a paradigm as a comprehensive perspective or framework used to understand a particular phenomenon. It includes a set of assumptions, theoretical models, and proposed solutions that guide how a subject matter is approached, what questions are asked, and how answers are interpreted. A paradigm shapes not only the way a researcher views reality but also influences the direction and methodology of the research itself.

In this study, the researcher adopts a post-positivist paradigm with quasi-qualitative approach. According to Creswell (2007), post-positivism is a paradigm that challenges and critiques the positivist doctrine regarding the nature of reality. It questions the assumption that reality can be fully understood through objective observation alone and emphasizes the influence of context, subjectivity, and the limitations of empirical inquiry.

Post-positivism shares some foundational principles with positivism, particularly in its commitment to empirical observation and logical reasoning. However, unlike positivism, post-positivism recognizes that absolute truth is ultimately unattainable. In this paradigm, researchers acknowledge that the

research process—especially during data collection and analysis—is often affected by various limitations, such as issues of data credibility, incompleteness, and interpretive bias. Therefore, the researcher cannot possibly obtain the absolute truth (Rahardjo, 2023).

Meanwhile, the quasi-qualitative approach is derived from the post-positivist paradigm. It resembles qualitative research in appearance and intent, but it is not entirely qualitative in nature. Quasi-qualitative is a research method that resembles qualitative research or appears to be so. This design is not a fully qualitative type, but rather a form of quantitative method that has been adapted to better reflect qualitative characteristics, hence the term “quasi.”

The concept of quasi-qualitative research emerges from the post-positivist paradigm, so it can be said to be an effort to incorporate qualitative elements into a quantitative approach. Thus, quasi-qualitative is defined as an approach that seeks to incorporate qualitative components within a basic quantitative research framework, although it does not fully reflect the characteristics of qualitative research (Rahardjo, 2023).

In applying the post-positivist paradigm, researchers formulate specific research questions. Analysis is conducted using thematic techniques based on selected YouTube video transcripts. Verbal interactions are categorized based on their function in constructing meaning, while visual and non-verbal elements—such as gestures, expressions, and spatial layout—are examined to understand their communicative significance. The chosen paradigm supports the use of both

qualitative and quantitative strategies, enabling a comprehensive examination of communication practices.

Post-positivism emphasizes the importance of methodological rigor, validity, and reliability, enabling researchers to approach the study with critical reflection and awareness of potential biases or limitations. Thus, the post-positivist paradigm provides a solid foundation for conducting detailed, reflective, and context-sensitive investigations into how multimodal communication is constructed and interpreted in digital media content.

B. Data Source

For data sources, researchers use the concept of cyber research as a data source. This concept uses social media as a reference for data collection. Researchers take data from conversations between members on a YouTube channel entitled “*Can We Guess This Celebrity Perfume? – Common Sense*”.

To clarify the video, it can be accessed at the link below:

<https://youtu.be/jgvyOHpKU0c?si=S-GCZOQI6YzR2Jma>

C. Data Analysis

This research was conducted through six stages of thematic analysis. The analysis followed the thematic analysis procedure by Braun & Clarke (2006), in the first phase, researchers familiarized themselves with the data by recording all field observations and transcribing verbal data from videos. The researcher also rewatched all data sources to ensure a deep understanding of the communication

context. In the second phase, the researcher began generating initial codes by organizing simple categories to facilitate pattern recognition in the data. These codes served as the basis for identifying emerging multimodal constructions.

The third phase was conducted by reading the entire transcript and beginning to look for communication patterns or categories of meaning that emerged consistently. After that, in the fourth phase, the researchers reviewed the identification results to select the most appropriate patterns. Because some parts of the data contained more than one form of multimodal communication, the researchers conducted an evaluation to determine the most dominant category. In the fifth phase, the researcher defined and named each of the main analysis units that had been formulated. Finally, in the sixth phase, the researcher compiled a report of the findings based on these categories and communication patterns to explain how verbal, visual, and non-verbal elements interacted in constructing meaning in The Try Guys videos.

D. Data Collection

In the context of this study, researchers act as “human instruments” because they are directly involved in all stages of implementation, from data collection to analysis. Researchers are responsible for selecting the appropriate methods, tools, and techniques to ensure that the data collection and analysis processes run effectively. In addition, researchers are also required to maintain data quality and interpret findings carefully in order to produce meaningful conclusions. According to Creswell (2009), researchers are the main instruments

in research because they are the ones who collect, process, and understand the data. This is in line with Rahardjo's (2020) opinion, which emphasizes that in a qualitative approach, researchers function as the main tools in the research process.

E. Research Instrument

The main instrument used in this research is the researcher herself, as the primary data interpreter and analyst. In the context of this study, researchers act as “human instruments” because they are directly involved in all stages of implementation, from data collection to analysis. Researchers are responsible for selecting the appropriate methods, tools, and techniques to ensure that the data collection and analysis processes run effectively. In addition, researchers are also required to maintain data quality and interpret findings carefully in order to produce meaningful conclusions. According to Creswell (2009), researchers are the main instruments in research because they are the ones who collect, process, and understand the data. This is in line with Rahardjo's (2020) opinion, which emphasizes that in a qualitative approach, researchers function as the main tools in the research process.

CHAPTER IV

FINDINGS AND DISCUSSION

This chapter contained the research findings and discussion. In the findings, the researcher focuses on the analyzed data. The discussion covered the findings of the data analysis.

A. FINDINGS

The findings presents of the study based on the analysis of multimodal elements found in the Try Guys YouTube video titled “Can We Guess This Celebrity Perfume?”.

Datum 1

Verbal Mode

The spoken utterance in this scene is:

“I’m gonna make these five contestants use each of their five senses like their life depends on it. But really it’s just to win a gift card.”

This verbal mode features a clear contrastive structure built through hyperbole and understatement. The phrase “like their life depends on it” is an exaggerated comparison, invoking a dramatic, high-stakes situation. Immediately afterward, “just to win a gift card” deflates the seriousness by inserting an anticlimactic payoff. This duality creates verbal irony, a hallmark of comedic discourse.

In terms of representational meaning (Kress & van Leeuwen, 2006), the utterance describes the goal of the contestants—engaging their five senses—thus presenting the content of the show. Interpersonally, the use of slang (“gonna”) and casual diction invites solidarity with the audience, establishing an informal and humorous tone. The language constructs a relationship between host and viewers that is playful and self-aware, aligning with the comedic genre of the program.

Visual Mode

During the delivery of this utterance, the host is seen with a wide smile and slightly raised eyebrows, conveying excitement and mock seriousness. His gaze is directed toward the camera, establishing direct contact with the audience. Following Kress & van Leeuwen’s (2006) framework, visual elements such as facial expression and gaze contribute to interpersonal meaning, shaping how the audience is positioned—not as distant observers, but as invited participants in the humor. The colorful, patterned background also supports the playful tone of the scene.

Gesture Mode

The host accompanies his speech with broad, open-handed gestures, extending his arms outward during the phrase “use each of their five senses”. This movement dramatizes the message and visually emphasizes the intensity of the challenge. According to multimodal theory, gestures serve to amplify verbal meaning and convey emotional tone. Here, the exaggerated hand movements

synchronize with the hyperbolic language, adding physical expressiveness to the verbal irony.

Audio Mode

The host's vocal delivery starts with a high, excited intonation on "I'm gonna make these five contestants...", then peaks dramatically at "like their life depends on it". In contrast, the phrase "just to win a gift card" is delivered with a deliberately flat tone and slower tempo. This modulation in pitch, rhythm, and volume is typical of audio mode in comedic performance. It cues the audience to recognize the shift from mock-serious to anticlimactic, enhancing the ironic and humorous function of the speech.

Spatial Mode

The host stands centrally in the frame, with balanced spacing between him and the colorful graphic background. His body faces directly forward, occupying frontal and prominent space in the visual composition. The contestants are yet to be introduced, keeping the host as the sole focus of attention. This spatial framing assigns the host a dominant narrative position, in line with his role as guide and entertainer. In multimodal terms, this mode contributes to ideational structure, reinforcing hierarchy and turn-taking in the scene.

Datum 2

Verbal Mode

The spoken utterance in this scene is:

Chris: "I'm Chris and my favorite sense is sound."

Gadiel: "I'm the king. The king of kings Gadiel."

Zach: "Everybody's lying if they don't think this is the most important one."

The contestants' verbal utterances reveal how language is used to perform identity, assert confidence, and introduce mild humor. Chris says, "I'm Chris and my favorite sense is sound," delivering a factual and neutral introduction. Gadiel follows with "I'm the king. The king of kings Gadiel," introducing metaphorical language and self-aggrandizement. Zach concludes with "Everybody's lying if they don't think this is the most important one," embedding humorous generalization and subjective assertion.

From a representational perspective (Kress & van Leeuwen, 2006), these utterances communicate identity positioning and sensory preferences, establishing who the speakers are and what they value. Interpersonally, Gadiel's statement is the most marked—his repetition and metaphor "king of kings" not only entertain but project dominance and charisma, suggesting performative masculinity and comic exaggeration. Zach's assertion, meanwhile, adopts a conversational style that indirectly engages the audience and implies playful judgment. These utterances reflect the informal, performative tone of the show, shaping a social bond with viewers

Visual Mode

Visually, each contestant is presented in a separate frame or segment, maintaining a medium close-up shot that captures facial expressions and upper body movements. Chris appears with a mild, relaxed smile, reflecting sincerity. Gadiel displays a bold facial expression, possibly lifting his chin slightly or raising his eyebrows to amplify his “king” persona. Zach’s expression includes raised brows and a cheeky grin, emphasizing the ironic tone of his verbal claim. In line with Kress & van Leeuwen’s (2006) theory, facial expressions and visual gaze here serve as cues for interpersonal meaning, guiding how the viewer is invited to interpret each contestant’s personality. While Chris maintains friendly neutrality, Gadiel establishes playful authority, and Zach evokes humorous skepticism. The visual mode thus complements and intensifies the linguistic content.

Gesture Mode

Each speaker employs gestures that reinforce their verbal identities. Chris may slightly point to himself or use minimal hand movement, signaling calm confidence. Gadiel uses assertive hand gestures—possibly placing his hand on his chest or making broad, king-like motions—to embody his metaphorical title. Zach might point or gesture outward, using mock-authoritative movements to match his satirical tone. According to multimodal theory, gesture mode functions as an extension of verbal expression, particularly in performative contexts. Gadiel’s gesturing supports the hyperbolic self-presentation, while Zach’s physical

mannerisms add comic rhythm and reinforcement to his verbal playfulness. The alignment of gesture and speech here creates embodied meaning, where physical movement becomes integral to discourse construction.

Audio Mode

Each contestant's tone of voice, pitch, and delivery pace convey attitude and performance. Chris speaks in a steady, even tone, reflecting straightforwardness. Gadiel uses a rising intonation with amplified volume, reinforcing his metaphorical self-positioning. Zach's speech contains a sarcastic inflection, with emphasized stress on "lying" and "most important", delivered in a knowing, comedic tone. These vocal features belong to the audio mode, which in Kress & van Leeuwen's multimodal system serves to deliver emotional coloration and rhythm to the utterance. The shifts in pitch and tempo signal how seriously or playfully the message should be interpreted. Gadiel's dramatic intonation elevates his character, while Zach's tone contributes to ironic detachment and comic effect.

Spatial Mode

Spatially, the contestants are presented in separate frames, standing individually against visually distinct backdrops or similar set designs. Each is centered in the screen, with the camera positioned at eye level, creating a sense of equality in framing. No one is spatially dominant in the composition, but their stance, frontal positioning, and body orientation toward the camera reinforce that they are speaking directly to the audience. In spatial semiotics, this use of frontal, symmetrical layout connotes direct address and narrative clarity. It supports the

ideational structure of this segment: introducing characters one by one in a neutral but performative environment. The spacing and orientation position the contestants not only as participants in the game, but also as entertainers performing to an unseen audience.

Datum 3

Verbal Mode

The spoken utterance in this scene is:

“Keith has a bomb!”

The verbal utterance in this scene includes the sudden exclamation “Keith has a bomb!” followed by the clarification, “A bath bomb.” The first statement activates a moment of ambiguity and tension, due to the polysemous word “bomb”, which typically connotes danger or threat. However, this tension is immediately dispelled by the second utterance, revealing that it refers to a harmless bath product, not an explosive device.

In terms of representational meaning, the language shifts from what seems like a serious accusation to an anticlimactic clarification, aligning with comedic timing and genre expectations. Interpersonally, the exaggerated language is not meant to inform but to provoke surprise and laughter. The contrast between the two utterances demonstrates verbal irony, as the meaning flips from alarming to absurd. According to Kress & van Leeuwen (2006), verbal mode in such cases is a

key tool for managing audience reaction and manipulating interpretation through pacing and contrast.

Visual Mode

Visually, the scene displays a rapid series of expressions. The initial exclamation is met with wide-eyed, shocked expressions from the other contestants. Moments later, once the clarification is made, these expressions shift to relief and laughter, including smiles, visible release of tension, and in some cases, exaggerated eye rolls or shoulder shrugs. Keith, holding the bath bomb, also displays a somewhat unbothered or confused facial expression, adding to the comedic contrast.

According to the visual metafunction in Kress & van Leeuwen's (2006) framework, facial expressions serve not just to reflect internal emotion but to cue viewers on how to interpret the scene. The transition in visual affect from fear to amusement guides the viewer from potential alarm to shared humor, making the shift visible, participatory, and multimodally anchored.

Gesture Mode

Gestures in this moment are critical to how meaning unfolds. Keith is seen holding up a round, colored object—the bath bomb—in a gesture that, out of context, may visually resemble holding something potentially dangerous. Other participants physically recoil, raise their hands, or move back slightly in initial reaction. Once the misunderstanding is cleared up, gestures shift to more relaxed forms—laughing, leaning forward, or casually pointing.

Gestures here amplify the ambiguity caused by the verbal mode and temporarily support the mistaken reading of the object. As multimodal discourse theory explains, gestures often disambiguate or reinforce the meaning of speech, but in this case, they intensify the confusion, making the comic reversal more satisfying when clarified. Thus, gesture mode contributes both to building tension and to resolving it.

Audio Mode

The audio mode is rich with dramatic fluctuation. The initial utterance “Keith has a bomb!” is delivered with an elevated pitch and urgency, mimicking panic or alarm. This is followed by quick overlapping vocal reactions from the other contestants—gasps, half-shouts, and scattered laughter. Once “A bath bomb” is clarified, the audio shifts to relief sounds and loud laughter, including background chuckling and sighs of release. Kress & van Leeuwen (2006) note that audio mode plays a crucial role in establishing emotional dynamics and rhythmic flow. Here, the dramatic audio crescendo serves to pull the viewer into the tension, while the release through laughter reinforces the comedic rhythm that is typical of this genre. The pacing and volume variations serve as audio cues that guide the audience’s affective experience.

Spatial Mode

Spatially, the scene includes Keith entering the shot or leaning into frame, holding the bath bomb, which creates a visual disruption—his body movement is what draws attention. The other participants are seated or standing at a distance,

and their immediate repositioning (e.g., moving away or reorienting their bodies) creates a momentary spatial shift, increasing tension and then resolving it as they relax back. In multimodal spatial design, such shifts are indicators of interpersonal dynamics and can signal moments of power, danger, or narrative focus. Keith becomes the spatial focus temporarily, and the camera may even pan or zoom slightly to emphasize his presence. After the misunderstanding is resolved, spatial relations return to normal, reflecting a re-stabilization of the scene.

Datum 4

Verbal Mode

The spoken utterance in this scene is:

“I said fresh laundry... and I drew the Snuggly Bear”

The utterance “I said fresh laundry... and I drew the Snuggly Bear” reflects a descriptive and associative use of language, rooted in sensory memory. The phrase “fresh laundry” evokes a specific olfactory experience, while the reference to “Snuggly Bear” (a familiar cartoon-like character) introduces a cultural and nostalgic association. From the lens of representational meaning (Kress & van Leeuwen, 2006), the verbal content constructs an imagined link between scent and memory—shaping an affective response that moves beyond literal identification. Interpersonally, the phrase is delivered in a confident yet humorous tone, suggesting playfulness and an intention to entertain as much as to guess accurately. It reflects how language in multimodal discourse can shift between

personal expression, imagination, and entertainment, especially in informal, humorous genres.

Visual Mode

In this moment, the visual mode includes the hand-drawn image of Snuggly Bear on the whiteboard or notepad, likely characterized by simple, cartoon-like lines. The image visually translates the contestant's verbal memory into a symbolic, childlike visual representation. This choice conveys not just a guess, but an attempt to link emotion, memory, and character into a concrete visual form.

According to Kress & van Leeuwen (2006), visual mode here operates as a semiotic translation of sensory experience. It externalizes an inner, olfactory perception into a shared visual reference. The use of a known figure (like Snuggly Bear) is not arbitrary—it activates cultural intertextuality, potentially shared by the audience, strengthening the interpersonal appeal.

Gesture Mode

The contestant is shown smelling the soap, often with dramatic or exaggerated movements—such as raising the bar close to the nose, closing the eyes, and taking a deep breath. These gestures are followed by pointing at the drawing or presenting the answer board to others, which serve to perform and validate their answer in a humorous, almost theatrical way. Gesture mode here is highly performative. According to multimodal discourse theory, gestures amplify the verbal message and visually frame the participant's thought process. The

sniffing gesture communicates engagement and concentration, while the presentation gesture asserts ownership of the (comedic) answer.

Audio Mode

The statement is delivered in a light, confident tone, with emphasis on “fresh laundry” that suggests personal certainty. The follow-up phrase, “and I drew the Snuggly Bear,” is often said with a slight laugh or smirk, marking the speaker’s awareness of the humorous nature of their answer. Audience members and fellow contestants often react with laughter, chuckles, or verbal affirmations. Audio mode here contributes to both emotional expression and group cohesion. Variations in tone, rhythm, and emphasis help position the speaker’s answer as both sincere and entertaining. These nuances—common in spoken discourse—guide the audience to recognize the humorous intention without it being explicitly stated.

Spatial Mode

Spatially, the contestants are positioned in a line or semi-circle, taking turns stepping forward or lifting their answer boards to share their guesses. The speaking contestant temporarily becomes the center of attention, framed more tightly by the camera while others wait. This temporary spatial focus indicates a shift in discursive authority—each contestant, for a moment, controls the scene. This dynamic turn-taking structure reflects hierarchical flattening, as each participant is given equal narrative space. According to Kress & van Leeuwen,

spatial design functions ideationally to show roles, shifts in focus, and interaction patterns.

Datum 5

Verbal Mode

The spoken utterance in this scene is:

“Kylie Jenner” → “Britney Spears” → “Kim Kardashian”

The spoken content of this scene is built around speculative guessing and humorous banter. Contestants call out names like “Kylie Jenner”, “Britney Spears”, and finally “Kim Kardashian.” These guesses are framed less as serious deductions and more as playful, performative statements, often based on social stereotypes or pop culture associations rather than scent profiles. From a representational perspective, the verbal mode reflects participants’ attempts to make sense of the smell through cultural references. Interpersonally, the verbal exchanges become a form of social bonding, where contestants build humor off each other's guesses. Some guesses are delivered with mock certainty, others with sarcasm or exaggeration, reflecting ironic detachment common in the show’s genre. According to Kress & van Leeuwen (2006), such verbal choices help establish tone, stance, and the speaker’s role in the discourse—here, as playful co-participants in a shared comedic game.

Visual Mode

Visually, the contestants' facial expressions shift dramatically as they smell the perfume—ranging from curious sniffing to sudden recoils, eye squints, or even mock-seductive looks. Some contestants may raise eyebrows, wrinkle their noses, or smirk, expressing both amusement and exaggerated emotional reaction. Meanwhile, Keith maintains a theatrical expression, often delivering the perfume with a playful or overdramatic flair. These visual cues are central to the interpersonal meaning, helping viewers understand not just what's happening, but how to emotionally respond. Facial expressions, in Kress & van Leeuwen's terms, act as visual semiotic resources that support the tone of the interaction. The camera often lingers on these expressions, emphasizing the comedic performance of the moment.

Gesture Mode

Gestures dominate this segment. Keith sprays perfume while leaning in dramatically, sometimes holding the bottle near a participant's face or neck. In one instance, a contestant leans in to sniff Keith directly, creating a comical intimacy. Other gestures include contestants waving their hands in front of their faces, grabbing their noses, or posing theatrically after giving a guess. These gestures serve both interactive and performative functions. They emphasize the sensory absurdity of the task (smelling someone else to guess a perfume) while also maintaining the humorous pacing of the scene. Gesture mode here bridges the

physical experience (smelling) and the social reaction (laughing, teasing), demonstrating the embodied dimension of multimodal discourse.

Audio Mode

The scene is underscored by light romantic or whimsical background music, enhancing the irony of the exaggerated intimacy. Contestants deliver their guesses in playfully uncertain tones, with rising intonations (“Kylie Jenner?”) or mock-dramatic declarations (“It’s Britney Spears. I can feel it.”). Audience laughter or off-screen chuckles (from the crew or other contestants) often punctuate the moment. In audio mode, tone of voice, musical score, and group laughter co-create the affective experience of the scene. According to Kress & van Leeuwen (2006), these sound elements help shape the emotional structure of the text—here, guiding viewers to interpret the scene as playful and exaggerated. Shifts in pitch and rhythm act as prosodic markers of irony, confusion, and surprise.

Spatial Mode

Keith moves between the contestants, positioning himself differently with each participant. The camera tracks his movement, creating a dynamic, shifting frame. Each contestant takes turns receiving the perfume spray, briefly becoming the spatial and narrative focus. The alternation between close-up shots and medium group shots highlights both individual reactions and collective interaction.

Datum 6

Verbal Mode

The spoken utterance in this scene is:

“Beans!” → “Ravioli!” → “Beef Stroganoff!”

The verbal expressions in this scene are short, spontaneous, and often delivered with uncertainty or comic exaggeration. Contestants make rapid guesses such as “Beans!”, “Ravioli!”, and “Beef Stroganoff!”, often with interjections like “Ugh!” or “Ew, what is this?” These utterances reflect immediate sensory interpretations and are shaped by the emotional reaction to the tactile experience. From a representational standpoint, these words name and interpret the physical texture of the hidden item. However, interpersonally, they function as part of the entertainment, creating humor through overreaction, absurdity, or surprise. According to Kress & van Leeuwen (2006), verbal mode here conveys not only content but also stance, evaluation, and affect, all of which are central in comedic performance.

Visual Mode

Visually, the scene is characterized by dramatic facial expressions, such as furrowed brows, grimaces, wide-open mouths, and other exaggerated signs of disgust, confusion, or surprise. Contestants’ faces often scrunch up in reaction to what they touch. When one says “Beef Stroganoff!”, their face may combine horror and humor. These facial expressions are key to the interpersonal

meaning—they guide viewers on how to emotionally interpret the moment. As Kress & van Leeuwen (2006) state, visual mode supports engagement and attitudinal positioning. Here, exaggerated expressions serve as visual punchlines to the physical action and verbal guesses, forming part of the show’s visual humor grammar.

Gesture Mode

Gesture plays a primary role in this scene. Contestants reach into a container blindly, groping, recoiling, and then re-reaching into the substance. Movements include tentative pokes, sudden jerks, or even shuddering motions of the hands and arms. These gestures reflect an interplay between caution and curiosity, performed for both sensory exploration and audience amusement. In multimodal discourse, gesture mode complements and sometimes even precedes verbal articulation. Contestants often move their hands before speaking, using touch as the first layer of meaning-making. Their full-body reactions—pulling away, stepping back, or shifting weight—function as embodied signs of affective engagement with the unknown object.

Audio Mode

Audio elements include exclamations of shock or revulsion, loud “ewws”, surprised gasps, and uncontrolled laughter from other contestants and possibly the production crew. The background music is playful, often timed to match reactions or intensify the tension right before the guessing. The phrase “Ravioli!” may be delivered in a higher pitch, rushed tempo, or with dramatic stress. Kress & van

Leeuwen (2006) suggest that audio mode contributes to the emotional rhythm and mood regulation of a scene. In this case, changes in tone, pitch, and volume signal escalating hilarity or uncertainty. The audio environment encourages viewers to share in the surprise and laughter, reinforcing the show's communal humor.

Spatial Mode

Spatially, each contestant stands or sits in front of a concealed container with their eyes covered or blindfolded, emphasizing the restriction of vision. Their posture leans forward, with their hands as the focal point of interaction. The camera typically zooms in on the hand-bucket interaction, or cuts to a split-screen view showing both the tactile moment and the contestant's face. This spatial composition reinforces the importance of physical contact as meaning-making. As Kress & van Leeuwen note, spatial layout reflects the ideational structure—what is being focused on and why. Here, the positioning of bodies, containers, and camera angles emphasizes the tactile mode as primary, making space itself a key participant in the communicative event.

Datum 7

Verbal Mode

The spoken utterance in this scene is:

“Moving on to the next round, let's go.” “Contestants, are you ready for the next round?”

The scene opens with verbal cues from the host: “Moving on to the next round, let's go.” He then calls out, “Contestants, are you ready for the next round?”, eliciting cheers. Follow-up lines such as “You're gonna put on those blindfolds” and “he is gonna bring some mysterious can of food and put it into your mouth” establish the rules of the next challenge in an informal, performative register. The verbal language here operates on multiple levels. Representationally, it introduces the task clearly. Interpersonally, phrases like “let's go” and “Open up them mouths” are casual, energetic, and mildly absurd, which reinforce the humor and informality of the show. The playful use of imperatives contributes to command-as-comedy, a recurring trope in Try Guys discourse. As per Kress & van Leeuwen (2006), verbal mode here not only transmits information but also establishes tone and power dynamics—Ryan and Keith momentarily take control as facilitators of chaos.

Visual Mode

Visually, the scene shows Keith approaching Zach while holding the can, with blindfolded participants visible, possibly smiling, laughing, or expressing mock concern. Their body language and facial expressions—even partially obscured—help communicate the humorous discomfort and anticipation. These visual elements contribute to interpersonal positioning. While the verbal commands come from Ryan, the visual focus shifts to Keith's theatrical entrance and Zach's vulnerable position as the first participant. The combination of blindfolds and open mouths also plays into comedic visual exaggeration—a familiar strategy in video entertainment. Kress & van Leeuwen's visual grammar

allows us to read the facial, bodily, and compositional cues as active participants in meaning-making, shaping how the viewer emotionally engages.

Gesture Mode

Keith's gestures are a central feature. He approaches Zach with deliberate body movement, possibly exaggerated strides or hunched shoulders, dramatizing his role as the "deliverer of the can." His gestures are embodied performance, further emphasized by his vocal improvisation (see Audio Mode). Gesture mode here includes: Keith's movement toward Zach, Ryan pointing or gesturing while giving instructions, Zach possibly flinching or preparing himself. As Kress & van Leeuwen argue, gestures extend and intensify spoken meaning. In this scene, gestures synchronize with the unfolding absurdity—enhancing comedic pacing and performativity.

Audio Mode

Audio is a standout mode in this scene. Keith's non-verbal vocalizing—a series of humorous sounds as he approaches ("vocalizing")—serves to build comedic tension, mimic suspense music, or simply entertain. Meanwhile, Ryan's energetic tone and contestants' cheering build an upbeat, ridiculous atmosphere. The contrast between Keith's silly sounds and Ryan's exaggerated directives (e.g., "Open up them mouths") emphasizes the role of intonation, rhythm, and vocal texture in humor. These audio cues signal how the viewer should interpret the scene—not as threatening or uncomfortable, but as absurd and performative.

According to Kress & van Leeuwen, audio mode sets the emotional flow of interaction, and here it reinforces chaos-as-comedy.

Spatial Mode

Spatially, Keith's movement toward Zach changes the scene's dynamic. The camera likely tracks this action, bringing Keith into the foreground. The other contestants remain seated or standing, blindfolded, creating a visually layered space with different participant roles. This shifting spatial layout—host standing, contestants seated or passive—signals a temporary reorganization of narrative control. Zach becomes the focal point, visually and narratively. The blindfolds create spatial tension between the seen and the unseen. According to Kress & van Leeuwen's framework, spatial arrangement reflects power, gaze, and focus—all of which are at play here, mediated comedically.

Datum 8

The spoken utterance in this scene is:

Is it a condom?" to "Hamburger," "Ox tail," "Corn dog," "Tacos?" and "Burrito."

Verbal Mode

The verbal mode in this scene is marked by rapid, spontaneous guessing, interspersed with humorous questions and reactive commentary. The guess "Is it a condom?" stands out as an absurd and unexpected association, instantly shifting the tone toward humor. Other guesses, such as "Ox tail," "Corn dog," and

“Tacos?”, show how participants try to reconcile tactile impressions with familiar food schemas. The host’s curt responses—“Wrong,” “Incorrect,”—act as control mechanisms within the game, while maintaining comedic rhythm. From a representational perspective, the verbal mode encodes participants’ attempts to interpret tactile data using linguistic categories. Interpersonally, the language use—especially absurd guesses and playful questions like “What does turmeric even taste like?”—functions to sustain audience engagement through relatability and humor. As Kress & van Leeuwen (2006) explain, verbal mode doesn’t merely convey ideas but also manages stance, tone, and relational dynamics—all of which are in constant flux throughout this scene.

Visual Mode

Though the transkrip doesn’t explicitly describe visual elements, the multimodal context allows us to infer a high degree of expressive facial reactions, especially in response to surprising textures or funny guesses. Contestants likely display furrowed brows, open mouths, raised eyebrows, or wide eyes, expressing confusion, disgust, or amusement. Visual mode contributes to interpersonal meaning by reinforcing the emotional content of the interaction. According to Kress & van Leeuwen, facial expression, eye movement, and head positioning all function as semiotic resources. In this scene, we can infer that participants’ visible struggle to identify the substance becomes part of the visual humor, communicated to the viewer not through words but through reaction shots and visual timing.

Gesture Mode

Gestures are central in this scene. Contestants insert their hands into the hidden containers, prodding, squeezing, or gently poking at the contents. Their body language may shift as they lean forward to explore more carefully, or jerk back in surprise or disgust. This exploratory gesturing is not only functional (to identify the object), but also performative—meant to signal uncertainty, discomfort, or confidence to others. As Kress & van Leeuwen (2006) argue, gesture mode is an important carrier of meaning, especially when verbal information is insufficient. In this context, gestures precede, parallel, or even replace speech, making them a primary meaning-making tool. The way hands move inside the container, or how quickly a participant recoils, helps shape the audience’s understanding of what is being felt, even without seeing the object.

Audio Mode

The soundscape of this scene is filled with collective laughter, quick interjections, and possibly a buzzer sound to indicate time running out or incorrect guesses. The delivery of guesses is often marked by rising intonation, stress, or elongated vowels (“Cooorn dooog?”), highlighting uncertainty or comic exaggeration. Audio mode here plays a dual role: (1) building tension as guesses escalate, and (2) releasing tension through shared laughter and sound effects. In Kress & van Leeuwen’s framework, the prosodic quality of speech—rhythm, stress, pitch—along with ambient sounds like laughter or buzzers, contribute to

the emotional architecture of the scene. These audio cues help the viewer interpret the moment not as competitive, but as comedic and collaborative.

Spatial Mode

In this scene, contestants are likely seated or standing in fixed positions before the containers, blindfolded to block vision. The framing emphasizes hands and faces, switching between close-ups of hand exploration and medium shots of full-body reactions. Each participant becomes the focal point during their turn, while others watch, laugh, or comment. Spatial arrangement reflects rotational authority—each contestant takes center stage in turn. The camera’s shifting focus marks this change, aligning with Kress & van Leeuwen’s ideational metafunction, which governs who or what gets narrative priority. Blindfolding removes visual dominance, forcing participants to rely on other senses and thereby increasing the salience of gesture, voice, and posture.

Datum 9

Verbal Mode

The spoken utterance in this scene is:

“Put on those blindfolds”, “You’re spelling DRAGON”

The verbal mode in this scene includes instructional language from the host (e.g., “Put on those blindfolds”, “You’re spelling DRAGON”), contestant complaints, and intentional verbal distraction from the “cluesers.” Participant lines such as “What the hell? None of these are what I need,” and “Okay, I got one

[bleep] letter” reveal emotional escalation, confusion, and the challenges of sensory deprivation. Gadiel’s line—“You know how to spell that huh? You don’t know how that is.”—functions as a verbal tactic of interference. From a representational standpoint, verbal mode communicates the game’s structure and participants’ real-time responses. Interpersonally, the use of slang, expletives, and sarcastic remarks heightens the intensity and humor of the moment. This aligns with Kress & van Leeuwen’s (2006) theory that verbal mode serves not only to convey content but also to enact social roles, express attitudes, and shape power dynamics—here, between focused players and disruptive “cluesers.”

Visual Mode

Visually, contestants are seen wearing blindfolds, moving through a space while searching for letters. Their body posture, direction of movement, and facial tension (when visible) contribute to the visual representation of struggle and concentration. Simultaneously, the “cluesers” may be visible hovering around or interfering, adding to the comedic disorder of the scene. In multimodal terms, the visual contradiction between the blindfolded contestants’ uncertainty and the cluesers’ confident interruptions adds dramatic and humorous tension. The unbalanced interaction is made visually clear through blocking, body proximity, and framing, all of which, per Kress & van Leeuwen, help construct interpersonal alignment and viewer positioning.

Gesture Mode

Gestures play a central role as contestants use tactile exploration to locate and identify letters without sight. Movements include cautious hand-sweeping, poking, grasping, and frustrated flinching or shrugging. These are nonverbal signs of disorientation and sensory compensation. Meanwhile, “cluesers” may employ gestures meant to confuse or mimic, such as pointing, waving hands near contestants, or moving letters around. These gestures contribute to both the narrative progression (finding letters) and the relational dynamic (competition versus sabotage). As noted in multimodal discourse theory, gestures become embodied extensions of thought and affect, conveying struggle, resistance, and interaction without needing verbal elaboration.

Audio Mode

The audio landscape is dense, with simultaneous dialogue, upbeat background music (starting at 17:48), and overlapping voices from contestants and cluesers. Contestant voices rise in pitch and volume as they express confusion or irritation. Interjections like “I can’t find—”, and background exclamations are accompanied by laughter and vocal chaos. Audio mode in this segment serves a rhythmic and affective function—building tension and comedy simultaneously. According to Kress & van Leeuwen (2006), rhythm, pitch, and sound layering provide crucial cues for emotional framing. The fast pace and overlapping audio signal urgency and confusion, while also reinforcing the scene’s light-hearted competitiveness.

Spatial Mode

The spatial configuration is complex and dynamic. Blindfolded contestants are positioned on the ground or in a designated search area, while “cluesers” circulate around them. This creates a contrast between stillness/focus (contestants) and mobility/disruption (cluesers). The camera likely follows both groups, alternating between wide shots (to capture chaos) and medium close-ups (to highlight expressions and hand movements).

Datum 10

The spoken utterance in this scene is:

“You’re going to have 10 cups and one guess to find that ball” and “Start the timer and 30 seconds go.”

Verbal Mode

The verbal mode in this scene is dominated by instructions, affirmations, and spontaneous reactions. The host sets the tone with direct and structured commands, such as: “You’re going to have 10 cups and one guess to find that ball” and “Start the timer and 30 seconds go.” These phrases establish the rules and time frame, representing the ideational metafunction of verbal mode. Meanwhile, contestant Ryann expresses confidence and performative self-assurance through lines like “I’m taking it home. This is my time.” and “I got this.”—delivered in a tone that suggests interpersonal positioning. As the round progresses, the verbal reactions shift to confusion and doubt: “Wow, it’s getting

really hard to follow,” “I lost it a bit in there,” and “I have no idea where it is.” These lines represent a narrative arc of psychological transition—from confidence to disorientation—and illustrate how verbal utterances convey not only content but also the changing emotional state of participants in real-time. Kress & van Leeuwen (2006) emphasize that verbal mode in multimodal discourse is deeply tied to attitude, stance, and power relations, all of which fluctuate here in seconds.

Visual Mode

Visually, this scene is marked by fast-paced hand movements, overlapping gestures, and tense facial expressions. The most visually dominant objects are the ten cups, quickly shuffled by multiple hands, which sometimes obscure each other, making tracking difficult. The fast interchanging hand positions form a visual blur that reflects the rising tension. The contestants’ facial expressions—such as furrowed brows, tightly focused eyes, or open mouths—suggest mental exertion and stress. The visual moment of selecting a cup becomes highly suspenseful, possibly emphasized by a camera close-up or framing zoom-in on the contestant’s hand. According to Kress & van Leeuwen, visual elements such as gaze, facial reaction, and object prominence are central in shaping interpersonal and narrative engagement. Here, the cups act as visual focal points, while contestant expressions regulate viewer empathy and anticipation.

Gesture Mode

Gesture plays a major role in this segment. The shufflers’ hands move rapidly, often crossing over one another to manipulate the cups, forming a

coordinated but chaotic gestural sequence. Meanwhile, contestants lean forward, track the motion with subtle head tilts, and exhibit physical stillness broken by sudden hand movements when choosing their final guess. Ryann’s vocal grunt as referenced in the transcript also indicates a bodily release of tension, often paired with micro-gestures like clenching fists or momentary hesitations before making a move. These gestures reflect embodied cognition—where meaning is formed through physical action as much as through speech or image. Gesture, in this context, is both strategic (tracking movement) and expressive (relaying internal state). In line with multimodal theory, it becomes a vital mode for conveying intention, uncertainty, and control.

Audio Mode

The audio landscape enhances the scene’s intensity. Background “intense music” begins at 17:48 and continues into this scene, synchronized with the speed of hand motion and building suspense. Layered above the music are distinct “cup clanking” sounds, adding realism and tempo to the challenge. Moments of verbal chaos—“everyone shouting”—combine with sharp statements (“One guess to win it all”) to orchestrate urgency. The combination of loud ambient sounds and sharp command lines defines the sound modality’s rhythmic structure, guiding how viewers emotionally experience the scene. Kress & van Leeuwen stress that sound modulation, background score, and vocal energy regulate the text’s emotional architecture. Here, the rising pitch of contestants’ voices and the increasing tempo of the background audio amplify the sense of imminent consequence.

Spatial Mode

The spatial arrangement in this challenge is highly orchestrated. Ten cups are placed in a straight line or clustered formation on a flat surface—likely a table. Contestants stand or lean directly in front of the setup, maintaining frontal orientation to keep their eyes fixed. The shufflers (possibly other players or crew) move around the table, temporarily invading the contestant’s visual field to create distraction. Keith’s line “Come up back here so the camera’s good for you” reflects an important spatial semiotic choice: optimizing the participant’s position for audience visibility, not just gameplay. This aligns with Kress & van Leeuwen’s spatial metafunction, where placement, proximity, and orientation convey not only relational roles but also narrative focus. Spatially, this scene organizes participants into roles—observer (contestant), performer (shuffler), and viewer (audience)—with camera work and body positioning reinforcing those distinctions.

B. DISCUSSIONS

The findings of this study strongly align with the multimodal discourse theory proposed by Kress and van Leeuwen (2006), which asserts that meaning is not constructed solely through language (verbal mode), but through the orchestrated use of multiple semiotic modes, including visual, gestural, audio, and spatial modes. These modes work together simultaneously, forming what Kress and van Leeuwen describe as a “multimodal ensemble.”

In the context of The Try Guys video, the five semiotic modes do not operate in isolation but are interdependently orchestrated to construct complex and layered meanings. Each datum in the analysis illustrates how these modes function dynamically in real-time to build humor, tension, interpersonal relationships, and audience engagement.

Verbal Mode and Interpersonal Meaning

According to Kress and van Leeuwen, verbal mode is not just representational but also serves interpersonal and textual functions. In the video, verbal communication is informal, highly expressive, and often humorous. For instance, phrases like “like their life depends on it” (Datum 1) or “I got one [bleep] letter” (Datum 7) are performative utterances that reflect emotion, exaggeration, and personal stance. These expressions do not merely inform—they construct tone, attitude, and interpersonal proximity with both fellow participants and the audience. The Try Guys' use of casual language, sarcasm, and irony underscores the interpersonal function of verbal discourse, as theorized by Kress and van Leeuwen.

Visual Mode and Audience Engagement

In multimodal theory, visual elements such as facial expressions, gaze, salience, and camera framing contribute significantly to meaning-making. This is clearly observed in moments where participants display exaggerated facial reactions—shock, disgust, amusement—especially during tactile guessing games (Datum 6 and 8). The use of close-ups, frontal shots, and expressive gestures

visually guides the viewer's interpretation and emotional alignment. Kress and van Leeuwen argue that visual grammar helps position the viewer, and this is evident as the Try Guys' eye contact with the camera and animated expressions foster a sense of direct engagement and inclusion for the audience.

Gesture Mode as Physical Meaning-Making

Gestures are not merely add-ons to speech but function as autonomous semiotic resources. In this study, gestures serve both expressive and representational purposes—such as reaching into containers, pointing at objects, flinching, or leaning. In situations where participants are blindfolded (Datum 7 and 9), gesture replaces sight and often even language as the primary mode of interpretation. This reflects Kress and van Leeuwen's claim that gestural communication becomes crucial in the absence of visual or verbal clarity, and can even dominate the meaning-making process in physical, embodied interactions.

Audio Mode and Rhythmic Structure

Kress and van Leeuwen emphasize that audio mode—including intonation, pitch, rhythm, ambient sound, and music—helps regulate the emotional tone and pacing of a text. In the Try Guys video, audio mode is used to heighten tension (e.g., intense music in Datum 10), to underscore humor (e.g., laughter tracks, bleeped expletives in Datum 7), and to enhance rhythm (e.g., synchronized music with dramatic gestures). These auditory features create a layered emotional texture, guiding viewers through suspense, relief, and

amusement. As theorized, sound is not neutral, but an active agent in shaping perception and engagement.

Spatial Mode and Social Dynamics

Spatial mode concerns the arrangement of participants and objects within a scene, their proximity, and their orientation. In the Try Guys video, spatial dynamics reflect the power structure and role distribution—such as contestants being blindfolded and seated while others walk around them to distract (Datum 9), or when a contestant moves into the frame to face the camera directly (Datum 10). Kress and van Leeuwen argue that space encodes social relationships, and this is evident as movement, camera angles, and bodily positions in the Try Guys' video visually represent shifts in focus, authority, and interactivity.

Through this discussion, the researcher aims to demonstrate that communication in The Try Guys video content is not solely constructed through language, but through a complex orchestration of semiotic modes that operate simultaneously and reinforce each other. The findings are discussed in relation to the two research questions, supported by the multimodal discourse theory of Kress and van Leeuwen (2006).

➤ Forms of Interaction Displayed by The Try Guys

Based on the analysis, members of The Try Guys exhibit various forms of social interaction that are layered, flexible, and performative in nature. They interact not only as contestants in a game, but also as entertainers, competitors, disruptors, and observers.

Competitive and Cooperative Interactions

Certain segments, such as Follow the Ball (Datum 10) and Touch and Guess (Datum 6 and 8), highlight competitive interactions where contestants vie to provide the correct answer and often assert their confidence with statements like “I’m taking it home. This is my time” or “I got this.” However, cooperative moments are also evident in the form of shared laughter, mutual encouragement, and playful commentary on one another’s actions. For instance, the collective laughter following the absurd guess “Is it a condom?” (Datum 8) reflects solidarity-based, socially driven humor.

Performative Interactions

In many moments, interactions are overtly performative—members intentionally construct personas or comedic identities. This is particularly visible during introductions (Datum 2), where Gadiel refers to himself as “the king of kings”, and Zach offers satirical commentary such as “Everybody’s lying if they don’t think this is the most important one.” These utterances are not merely self-expressions but are deliberate performances aimed at the audience, framing identity through multimodal performance.

Distractive Interactions

Another significant interactional type is distraction, particularly evident in the blindfold spelling challenge (Datum 9), where “cluesers” attempt to divert the contestants’ attention with misleading comments like “You know how to spell that huh?” or through creating chaotic noise. This introduces a unique power

dynamic and playful disruption into the game’s structure, producing deliberate comedic tension. In summary, the forms of interaction presented in the Try Guys’ content are highly fluid and context-sensitive—ranging from serious to absurd, cooperative to satirical—all of which are enriched by multimodal performativity.

➤ **Members “The Try Guys” use of Multimodal Elements in Building Communication**

Multimodality is not merely present in the Try Guys content—it is fundamental to how communication occurs. Through five semiotic modes—verbal, visual, gesture, audio, and spatial—meaning is constructed not only by what is said, but by how it is spoken, displayed, enacted, and positioned in time and space.

Verbal Mode: Informal, Spontaneous, and Dramatic

The verbal mode features casual, conversational language, including hyperbole, understatement, metaphor, and spontaneous interjections. Statements like “like their life depends on it” (Datum 1) or “I got one [bleep] letter” (Datum 7) demonstrate how language is used not just to convey information, but to establish tone, character, and comedic effect. The verbal mode is central in constructing interpersonal engagement and providing a framework for the game’s structure.

Visual Mode: Emotional Framing and Expressivity

The visual mode includes facial expressions, gaze direction, body posture, and vibrant background design. Moments of visible disgust, surprise, laughter, or confusion (e.g., Datum 6 and 10) visually signify emotional responses that support the verbal content. The use of close-up framing often highlights hands or facial expressions, enhancing emotional proximity between participants and viewers.

Gesture Mode: Embodied Meaning-Making

Gestures are heavily employed, including pointing, touching objects, raising hands, and spontaneous body reactions such as recoiling in disgust or leaning in with curiosity. In blindfolded situations (Datum 7 and 9), gesture becomes the primary communicative tool, central to exploration and interpretation. These movements are not only expressive but often replace language entirely, indicating how gesture operates as an autonomous mode of communication.

Audio Mode: Rhythm, Humor, and Tension

The audio mode consists of vocal intonation, volume, sound effects like buzzers, and background music. Intense music accompanying high-stakes gameplay (Datum 10), or censor bleeps during expletives (Datum 7), highlight how audio is used to regulate emotional pacing. Audio is not neutral; it is crafted to enhance suspense, signal humor, and create rhythm.

Spatial Mode: Body Positioning and Camera Dynamics

Spatial mode concerns the physical arrangement of participants and how the camera captures those interactions. In scenes such as blindfold spelling (Datum 9), spatial positioning delineates who is the narrative focus and who plays the disruptor. Control over space becomes a means of managing narrative flow and social hierarchy. Even directives like “Come up back here so the camera’s good for you” reflect a consideration for audience visibility and perspective, making spatiality both functional and semiotic.

The findings affirm that meaning in Try Guys content is not constructed by a single mode, but through the simultaneous coordination of multiple modes. A statement becomes humorous not merely due to its content, but through its intonation, facial expression, dramatic gesture, and collective laughter. Each scene thus becomes a multimodal performance, where participants are consciously performing not just for one another, but for the audience watching.

Kress and van Leeuwen (2006) describe this as orchestrated semiosis, in which every mode contributes a distinct but interdependent function. The Try Guys demonstrate a high level of competence in deploying all five modes—not just for clarity, but to maximize comedic timing, emotional impact, and interactive viewer engagement.

The data suggest that The Try Guys content is not merely a game show or variety piece—it is a highly curated example of multimodal meaning-making,

where performance, narrative, humor, and interaction are all produced through coordinated semiotic strategies.

CHAPTER V

CONCLUSION AND SUGGESTION

This chapter presents the conclusions drawn from the study on how the members of The Try Guys construct communication through multimodal elements. Using the framework of Kress and van Leeuwen's (2006) Multimodal Discourse Analysis, this chapter summarizes the key findings, addresses the research questions, and outlines the theoretical implications of the study.

A. Conclusion

This research was conducted to explore how the members of The Try Guys utilize various multimodal elements—namely verbal, visual, gestural, audio, and spatial modes—to build communication and create meaning in their YouTube video titled “Can We Guess This Celebrity Perfume?”. By applying the Multimodal Discourse Analysis framework by Kress and van Leeuwen (2006), the study focused on two main aspects: the forms of interaction among the members, and how these multimodal resources are strategically employed to engage the audience.

Based on the analysis, it was found that the interactions in the video are not solely reliant on spoken language. Instead, they are constructed through the collaboration of multiple modes that appear simultaneously. Each spoken utterance by the members is supported by facial expressions, body movements, vocal intonation, spatial positioning, and camera direction. This combination results in communication that is rich, expressive, and entertaining.

The interactions show that communication in the video is not just about solving the game or challenge, but also about creating a shared experience and collective entertainment. The elements do not function in isolation. Meaning is constructed through the interplay of all these modes. For example, a humorous line may be delivered with sarcastic wording (verbal), a high-pitched tone (audio), a mocking facial expression (visual), exaggerated gestures (gestural), and a well-framed camera close-up (spatial)—all of which contribute to the comedic effect.

The findings align closely with the theory proposed by Kress and van Leeuwen (2006), which states that communication is inherently multimodal, and each mode serves a specific function: ideational (what is being said), interpersonal (how relationships are formed), and textual (how the message is organized).

In The Try Guys video, the interpersonal function stands out prominently. The members do not only speak to each other, but also perform as if they are speaking directly to the viewers. Their delivery is crafted to feel casual and engaging, showing their awareness that their communication is being visually and emotionally consumed by an audience.

This highlights how YouTube content is a hybrid form of communication—part casual conversation, part performance. On one hand, it feels like a group of friends playing a game. On the other, it is a carefully staged performance for the camera. This is where multimodality becomes central: every visible and audible detail contributes to a layered performance of meaning.

In conclusion, this study finds that communication within The Try Guys video relies heavily on the combination of multimodal elements. No single mode operates alone; rather, they interact and complement each other. Through a blend of speech, facial expressions, vocal tone, gestures, and visual framing, the group creates communication that not only conveys information but also builds relationships, expresses emotion, and entertains effectively.

Furthermore, this study demonstrates that multimodal discourse analysis is a highly relevant approach for understanding how communication works in modern digital contexts, especially on interactive platforms like YouTube. Such analysis allows us to see that today's communication is a fully embodied performance, where everything that is seen and heard contributes meaningfully to the message being delivered.

B. Suggestion

This research opens opportunities for future researchers to expand the object of study to other digital platforms such as TikTok or Instagram Reels. These platforms possess unique characteristics and communicative dynamics that differ significantly from YouTube, particularly in terms of duration, content format, and interaction style. By exploring multimodal interactions across various media environments, future studies may uncover comparative insights on how meaning is constructed differently in each platform's specific context. Such comparative analysis would enhance our understanding of digital communication practices in diverse technological and social settings.

Moreover, the findings of this research can serve as practical guidelines for digital content creators. Understanding how verbal, visual, and gestural elements work together to shape messages and create emotional resonance with audiences can inform more intentional content strategies. Creators who apply multimodal elements in a balanced and deliberate way are more likely to produce engaging, relatable, and meaningful content. This knowledge can be especially valuable in designing interactive media experiences that deepen viewer connection and increase audience retention in the digital landscape.

For academics and scholars, this study contributes significantly to the development of contemporary discourse analysis, particularly within the realm of applied linguistics. By incorporating visual and audio modes into linguistic analysis, researchers are equipped with a more holistic framework for understanding how meaning is conveyed in digital media. Multimodal discourse analysis thus offers an adaptive and comprehensive methodological approach for studying the complexity of modern communication across multimedia platforms.

BIBLIOGRAPHY

- Benson, P. (2005). *YouTube as text Spoken interaction analysis and digital discourse*. <https://doi.org/10.4324/9781315726465-6>
- Bezemer, J., & Kress, G. (2015). *Multimodality, learning and communication: A social semiotic frame*. Routledge.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Bull, G., & Anstey, M. (2018). *The Literacy Labyrinth: Second Edition*. Pearson Australia.
- Burgess, J., & Green, J. (2018). *YouTube: Online video and participatory culture (2nd ed.)*. Polity Press.
- Chakma, K., Begum, U., & Das, S. (2022). *Heliyon YouTube as an information source of floating agriculture : analysis of Bengali language contents quality and viewers ' interaction*. *Heliyon*, 8(9), e10719. <https://doi.org/10.1016/j.heliyon.2022.e10719>
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches (2nd ed.)*. Sage Publications.
- Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches (3rd ed.)*. Sage Publications.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2011). *The SAGE Handbook of Qualitative Research (4th ed.)*. SAGE Publications.
- Fairclough, N. (2003). *Analyzing Discourse: Textual Analysis for Social Research*. Routledge.
- Fang, I. E. (2019). *Media literacy: A crash course in 101 ways to understand and critique the media*. Peter Lang Publishing.
- Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.
- Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar (4th ed.)*. Oxon: Routledge.
- Henry Jenkins, Sam Ford, and Joshua Green. *Spreadable media : creating value*

- and meaning in a networked culture, (New York University Press, 2013)
- Iedema, R. (2003). *Multimodality, resemiotization: Extending the analysis of discourse as multi-semiotic practice*. *Visual Communication*, 2(1), 29–57.
- Janvier, A. J., Pendleton, E. G., Mortensen, L. J., Green, D. C., Henstock, J. R., & Canty-Laird, E. G. (2022). *Multimodal analysis of the differential effects of cyclic strain on collagen isoform composition, fibril architecture and biomechanics of tissue engineered tendon*. *Journal of Tissue Engineering*, 13. <https://doi.org/10.1177/20417314221130486>
- Jenkins, H., Ford, S., & Green, J. (2013). *Spreadable media: Creating value and meaning in a networked culture*. NYU Press.
- Jewitt, C., & Jones, R. H. (n.d.). *The Routledge Handbook of Multimodal Analysis*. Routledge.
- Kadwa, M. S., & Alshenqeeti, H. (2020). International Journal of Linguistics, Literature and Translation (IJLLT) The Impact of Students' Proficiency in English on Science Courses in a Foundation Year Program. *International Journal of Linguistics, Literature and Translation (IJLLT)*, 3(11), 55–67. <https://doi.org/10.32996/ijllt>
- Kress, G., & Van Leeuwen, T. (1996). *Reading Images: The Grammar of Visual Design*. Routledge.
- Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Arnold Publishers.
- Kress, G. (2009). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.
- Lange, P. G. (2007). *Publicly private and privately public: Social networking on YouTube*. *Journal of Computer-Mediated Communication*.
- O'Halloran, K. L. (2016). *Multimodal analysis and digital technology*. In O'Halloran, K. L., Tan, S., & Smith, B. A. (Eds.), *Multimodal studies: Exploring issues and domains*. Routledge.
- Rahardjo, Mudjia. (2018). *Studi Teks dalam Penelitian Kualitatif (PDF)*. Disampaikan pada mata kuliah Metodologi Penelitian, Sekolah Pascasarjana Universitas Islam Negeri Maulana Malik Ibrahim Malang. (Unpublished)
- Rahardjo, M. (2020). *Qualitative Research Methodology for the Social and Sciences Humanities (From Theory to Practice)*. Malang: Republic of Media

- Rahardjo, M. (2023) What is Quasi-qualitative?. Delivered in the course Research Methodology, English Literature / Humanities, Odd Semester. (Unpublished)
- Rahardjo, M. (2025). Social Research Methodology: *Paradigms, Approaches, Methods and Techniques*. Presented at Social Research Intership in The Perspective of a Qualitative Approach, Merdeka University, 11 January 2025.
- Suparno, D., Thamrin, M. H., & Chairul, A. I. (2022). *Pengantar Multimodalitas dan Transitivitas*. Universitas Islam Negeri Syarif Hidayatullah Jakarta.
- Ulati, N. M. S. (2021). Multimodal Analysis of “ENERGEN” Ads. *International Journal of Systemic Functional Linguistics*, 4(1), 25–28.
- Walkington, C., Nathan, M. J., Huang, W., Hunnicutt, J., & Washington, J. (2024). Multimodal analysis of interaction data from embodied education technologies. *Educational Technology Research and Development*, 72(5), 2565–2584. <https://doi.org/10.1007/s11423-023-10254-9>
- Wilson, A., Wilkes, S., Teramoto, Y., & Hale, S. (2023). Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*, 10(12). <https://doi.org/10.1098/rsos.230964>

CURRICULUM VITAE



Fila Izul Mya was born in Malang on October 9, 2002. She graduated from Wahidiyah Senior High School Kepanjen in 2021. After completing her secondary education, she continued her academic journey at the State Islamic University of Maulana Malik Ibrahim Malang in the same year. She became a student of the Faculty of Humanities, majoring in English Literature, and successfully completed her studies in 2025. During her time at the university, she actively participated in several campus activities. She joined the Volunteer Corps (KSR PMI) of UIN Malang, an intra-campus organization dedicated to humanitarian work. In addition, she also took part in *Pesona Humaniora*, one of the student organizations within the Faculty of Humanities, as a means of academic and personal development.