Appendix

A Experimental Procedures

The experimental process began with an email from BCG's global CEO to all individual-level contributors within the organization, encouraging participation in a study aimed at assessing the impact of Generative AI on BCG's consulting work. This email, refined by the authorship team, included a link to an enrollment survey. A total of 852 individuals expressed their interest by responding to the survey, completing a short psychological assessment, and providing demographic information such as gender, native language, and educational background. This demographic information was subsequently used for stratified random assignment.

Approximately one month following the survey, respondents were invited to undertake their designated experimental tasks. The experiment required participants to allocate five consecutive hours for task completion (breaks were built in the experimental flow) within a window spanning five weeks between May and June 2023.[27] Of those who expressed interest, 758 consultants participated in the experiment: 385 were assigned to the inside-the-frontier tasks, and 373 to tasks situated outside-the-frontier. Appendix B describes the two sets of experimental tasks.

Participant engagement and effort were encouraged through performance-based incentives. All participants who fully engaged in the experiment were acknowledged with an "office contribution" recognition, which had financial implications for their annual bonuses. Additionally, outstanding performance was further recognized, with the top 20% of participants receiving special recognition and the top 5% receiving a small gift. Notably, this recognition was communicated to the committee responsible for overseeing participants' career progression and performance evaluations. In their interviews, participants confirmed these incentives were significant to them.

Two sets of evaluators were employed: MBA students with prior business experience assessed the inside-the-frontier tasks using rubrics detailed in Appendix C, which were tailored to evaluate creativity, analytical thinking, writing proficiency, and persuasiveness[28]; BCG consultants evaluated the outside-the-frontier tasks, using correctness (a binary variable, taking value 1 if subjects provided an accurate recommendation, and 0 otherwise) as the primary criterion. BCG consultants and MBA students with prior business experience also evaluated the subjective coherence quality of the recommendation in the outside-the-frontier task.[29]

Appendix C reports the rubrics human graders used for the inside-the-frontier task, based on the type of question they assessed (creativity, analytical thinking, writing proficiency, and persuasiveness). These assessments were additionally complemented

[27] All participants completed their work within the allotted timeline: the system prevented any extension and automatically captured responses at the designated endpoint. Performance evaluations were based exclusively on outputs generated within this standardized time frame.

[28] Separately, evaluations from product-design specialists replicated our results (see Appendix 5.2)

[29] One observation is missing.

(for robustness) by scores generated using GPT-4, as discussed in the main paper. Appendix D reports the GPT-4 prompt used to generate these grades. For the task outside the frontier, participants' answers were scored as correct if they responded with the correct response to the business case.[30] Appendix E reports the rubric that human graders used to evaluate the subjective coherence quality of the recommendation in the outside-the-frontier task.

Additionally, inside-frontier answers were scored by MBA students with prior business experience and, . Outside-frontier answers were evaluated by experienced BCG consultants and MBA students with prior business experience.

[30]Both BCG consultants and GPT-4 assessed these responses, agreeing in over 95% of cases. Any discrepancies were resolved by an independent human grader.

B Tasks

In the following Appendix, we detail the specific tasks undertaken by subjects during the experiments. Both experiments started with an assessment task, serving as the initial phase where all subjects were required to complete it without any assistance from Generative AI. This initial task is crucial as it establishes a baseline, capturing the abilities and skills of the subjects in the absence of AI support. The subsequent task represents the core experimental phase. Here, subjects have the opportunity to leverage Generative AI, with the level of access determined by their respective treatment assignments. We report these tasks below for both experiments.

Outside the Frontier

Both assessment and experimental tasks were designed with a maximum duration of 60 minutes; however, participants had the option to finish it sooner and advance to the subsequent phases. The prompts and questions for the tasks are provided below.

Assessment Task

The CEO, Harold Van Muylders, would like to understand which of the three distribution channels that the company uses (fully owned stores, franchisee stores, or online) to focus his efforts. Please find attached interviews from company insiders on this issue. In addition, the attached Excel sheet provides financial data broken down by distribution channels.

Aim: Please prepare a 500-750 word note to the CEO. The note should focus on the following:

- If the CEO must pick one distribution channel to focus on to drive profit growth in the company, what channel should that be? What is the rationale for this choice? Please support your views with data and/or interview quotations as appropriate.
- Please also suggest innovative and tactical actions the CEO can take to boost profit growth in your chosen distribution channel. Please be creative, and feel free to rely on your own business judgement on what is appropriate for Kleding.

Experimental Task

The CEO, Harold Van Muylders, would like to understand Kleding's performance by the company's three brands (Kleding Man, Kleding Woman, and Kleding Kids) to uncover deeper issues. Please find attached interviews from company insiders on this issue. In addition, the attached excel sheet provides financial data broken down by brands.

Aim: Please prepare a 500-750 word note to the CEO. The note should focus on the following:

- If the CEO must pick one brand to focus on and invest to drive revenue growth in the company, what brand should that be? What is the rationale for this choice? Please support your views with data and/or interview quotations.

- Please also suggest innovative and tactical actions the CEO can take to improve this chosen brand. Please be creative, and feel free to rely on your own business judgement on what is appropriate for Kleding.

Inside the Frontier

The assessment task was allocated a duration of 30 minutes, and the experimental task was set for 90 minutes. Subjects were required to use the full allotted time for each task and were not permitted to finish early. The prompts and questions for each task are detailed below.

Assessment Task

You are working for a beverage company in the unit developing new products. Your boss asked you to present an idea for a new product at the next manager meeting. Please, respond to the questions below.

1. Generate ideas for a new drink in markets that are underserved. Be creative, and give at least 10 ideas.
2. Pick the best idea, and explain why, so that your boss and other managers can understand your thinking.
3. Describe a potential prototype drink in vivid detail in one paragraph (3-4 sentences).
4. Come up with a list of steps needed to launch the product. Be concise but comprehensive.
5. Come up with a name for the product: consider at least 4 names, write them down, and explain the one you picked.

Experimental Task

You are working for a footwear company in the unit developing new products. Your boss asked you to present an idea for a new product at the next manager's meetings. Please, respond to the questions below.

1. Generate ideas for a new shoe aimed at a specific market or sport that is underserved. Be creative, and give at least 10 ideas.
2. Pick the best idea, and explain why, so that your boss and other managers can understand your thinking.
3. Describe a potential prototype shoe in vivid detail in one paragraph (3-4 sentences).
4. Come up with a list of steps needed to launch the product. Be concise but comprehensive.

5. Come up with a name for the product: consider at least 4 names, write them down, and explain the one you picked.
6. Use your best knowledge to segment the footwear industry market by users. Keep it general, and do not focus yet on your specific target and customer groups.
7. List the initial segments might you consider (do not consider more than 3).
8. List the presumed needs of each of these segment. Explain your assessment.
9. Decide which segment is most important. Explain your assessment.
10. Come up with a marketing slogan for each of the segments you are targeting.
11. Suggest three ways of testing whether your marketing slogan works well with the customers you have identified.
12. Write a 500-word memo to your boss explaining your findings.
13. Your boss would like to test the idea with a focus group. Please, describe who you would bring into this focus group.
14. Suggest 5 questions you would ask the people in the focus group.

    Now, imagine your new product entering the market.

15. List (potential) competitor shoe companies in this space.
16. Explain the reasons your product would win this competition in an inspirational memo to employees.
17. Write marketing copy for a press release.
18. Please, synthesize the insights you have gained from the previous questions and create an outline for a Harvard Business Review-style article of approximately 2,500 words. In this article, your goal should be to describe your process end-to-end so that it serves as a guide for practitioners in the footwear industry looking to develop a new shoe. Specifically, in this article, please describe your process for developing the new product, from initial brainstorming to final selection, prototyping, market segmentation, and marketing strategies. Please also include headings, subheadings, and a clear structure for your article, which will guide the reader through your product development journey and emphasize the key takeaways from your experience. Please also share lessons learned and best practices for product development in the footwear industry so that your article serves as a valuable resource for professionals in this field.

'Timing' is defined differently across tasks: for the inside-the-frontier arm we record seconds elapsed until participants reached the final screen (Q18), after which they waited out the fixed session clock, whereas for the outside-the-frontier arm we record total work-time because subjects could submit whenever their recommendation was complete.

C Evaluation Rubric - Inside the Frontier

In the evaluation process, human graders were instructed to carefully review the provided rubrics for each criterion and assess every response accordingly. We stressed the importance of focusing on the quality of responses rather than their length or any other attributes.

Each response was assessed using a specific rubric, which focused of the question type, whether it be creativity, analytical thinking, writing proficiency, or persuasiveness. The details of the criteria and scoring scales used in these rubrics are outlined below.

Creativity

- 1-2: The ideas are derivative or lack originality. No evidence of fresh thinking or new perspective.
- 3-4: Shows some original thought, but primarily builds upon existing ideas. Limited evidence of innovation or pushing boundaries.
- 5-6: There is a clear demonstration of some novel ideas and perspectives. These ideas show some degree of out-of-the-box thinking.
- 7-8: Shows a high level of originality and creativity. Ideas are innovative, unique, and they challenge the norm.
- 9-10: Exceptional creativity. Ideas are not only original and innovative, but they also redefine the norm and could potentially lead to significant shifts in understanding or practice.

Analytical Thinking

- 1-2: Does not clearly identify or understand problems or factors. Does not evaluate information effectively or identify solutions.
- 3-4: Identifies some problems or factors, but analysis is superficial or incorrect. Evaluates information and identifies solutions inconsistently or ineffectively.
- 5-6: Identifies most problems or factors correctly. Evaluates information and identifies solutions adequately, but there are some missed opportunities.
- 7-8: Identifies problems or factors accurately and comprehensively. Evaluates information effectively and identifies feasible and effective solutions.
- 9-10: Identifies problems or factors accurately, comprehensively, and insightfully. Evaluates information highly effectively and identifies creative, feasible, and highly effective solutions.

Writing Proficiency

- 1-2: Writing is unclear or confusing. Has significant issues with grammar, punctuation, and/or spelling. Lacks clarity and organization.
- 3-4: Writing is somewhat clear, but there are issues with grammar, punctuation, and/or spelling. Organization could be improved.
- 5-6: Writing is mostly clear and organized, but there may be minor issues with grammar, punctuation, and/or spelling. Organization is mostly good.
- 7-8: Writing is clear and well-organized. There are few, if any, issues with grammar, punctuation, and spelling.
- 9-10: Writing is exceptionally clear, well-organized, and concise. There are no noticeable issues with grammar, punctuation, and spelling.

Persuasiveness

- 1-2: Arguments are unclear or confusing. Does not provide compelling reasons or use evidence effectively to support points.
- 3-4: Some arguments are clear and logical, but many are not. Evidence used inconsistently or ineffectively to support points.
- 5-6: Most arguments are clear and logical. Uses evidence to support points, but occasionally the evidence does not strongly support the argument.
- 7-8: Arguments are clear, logical, and persuasive. Uses evidence effectively to support points.
- 9-10: Arguments are extremely clear, logical, and persuasive. Uses evidence highly effectively to support points. The writing could sway a neutral or even somewhat skeptical audience.

D GPT-4 Evaluation Prompt

We evaluated all assignments using GPT-4. The specific prompt employed for these evaluations is detailed below. Our results have been tested for robustness with various alternative prompts.

Imagine you are a writing professor at a renowned university. You have given your college students the following assignment below. You are a harsh grader and not afraid to give low scores (1, 2, or 3) if it is needed. But you are also fair, and recognize great quality to give high scores if needed (8, 9, or 10). Please, score the following assignments on a 1-10 scale, focusing on (creativity, analytical thinking, writing proficiency, persuasiveness). Please answer with just the score without any explanation or comments for the overall answer.

E Evaluation Rubric - Subjective Coherence Quality

Score	Description	Example
1	• Participant does not identify tactical actions for client to boost profits	*"From a prospective target of driving profit growth, the recommendation would be to further develop franchisee which is the channel with the strongest financials: very steady margins: ~2% CAGR for the full channel over the last 4 years, in line with revenue growth for the channel. The business model is very straightforward for the client with limited financial risks.* *In addition, there are very good prospects for the channel, as per News Article." (1)*
2-4	• Participant alludes to recommendations on how to boost profit but does not explicitly call out tactical actions. • Description of actions lacks specificity. Little to no description of business reasoning and impact on profit.	*"3. Gives time for Kleding to improve its fully owned and online channels.* *By shifting focus to a profitable channel, it will give Kleding time to fix its fully owned and online channels, which require more work. For fully owned stores, we need to evaluate the entire store portfolio and potentially close those that are unprofitable.* *We believe it is important to keep some so that we don't become overly reliant on franchisees and lose leverage in profit sharing negotiations. Meanwhile, we may need to hire people who are more knowledgeable about e-commerce to improve both the top-line and bottom-line." (3)*
5	• Participant identifies tactical actions for client to boost profits. • Tactical actions are aligned with overall channel strategy • Participant does not describe how to implement strategy. Explanation lacks specificity. • Business reasoning is unclear.	*"Going forward Kleding should deploy 3 tactics to increase profitability though franchisee stores* *1. Negotiate real time access to Kleding sales data at Franchise stores and the ability to marketing and collection team to make adjustments to visual merchandising and discounts based on the data* *2. A clear communication route between the Kleding team and Franchise store managers so changes in strategy can be applied immediately* *3. Training to Franchise store staff on customer engagement, product knowledge and tactics for competitive selling" (5)*
6-8	• Participant identifies tactical actions for client to boost profit • Tactical actions are aligned with overall channel strategy • Participant describes tactical actions in detail and outlines how to implement • Business reasoning lacks clarity. Participant does not fully connect	*"Key takeaway 3: Consider closing or shrinking underperforming stores* *The immediate action plan is to bring back the fully owned stores' strategy to the smaller/more intimate model. It's important to consider that some of the underperforming stores will be under long-term contracts, which can lead to fines and other immediate operational costs. Therefore, for some of the stores we need to consider whether it*

	explanation to client concerns. Impact on profit unclear.	*would cost more to close the stores or to keep them and have a negative contribution margin for the coming years. Trying to shrink the stores, instead of fully closing them can also lead to a more friendly contract renegotiation." (7)*
9-10	• Participant identifies tactical actions for client to to boost profit • Tactical actions are aligned with overall channel strategy • Participant includes elaborate description of how to implement tactical actions • Actions backed up by sound business reasoning. Participant draws on historical pain points and belief audits. • Participant outlines impact of action on profit. • Points deducted if explanation is incomplete	*"Tactical actions:* 1. *Close large unprofitable stores: as we learnt from Head of Finance, bigger stores mean higher rents, higher staff costs, and high amortization costs. Since none of the new stores opened since 2015 has been profitable, we should close those stores. Despite near term cost, it would help to stop losing money earlier than later.* 2. *Focus on quality over quantity: instead of showcasing all items in the collection, the stores should focus on selling the top items that have potential to be bestsellers. At the same time, it should try to create a more intimate shopping experience in influencing customers' decisions.* 3. *Invest in employee training programs: as Head of Owned Stores mentioned, the rapid expansion has led to lack of employee training, therefore less knowledgeable staff and suboptimal sales skills. By investing in employee training, they can understand the products deeper and target potential customers better, as well as improving their sales skills and customer service skills. With happier customers, there will be higher sales from returning customers, and overall brand image improvement, driving long term sales. There are also generative AI created sales simulations sellers can utilize to practice those skills.* 4. *Invest in advanced supply chain system: delays and stockouts can all lower customer satisfaction and lead to lower sales. With GenAI, there are more tools that can efficiently manage operations and provide customized suggestions."*

F Management Consultants as Knowledge Workers

Our experiment examines various tasks that collectively form part of a management consultant's job. These tasks differ substantially from one another. Many occupations similarly require workers to perform a range of distinct activities. For instance, at a broad level, college faculty typically teach, conduct research, and engage in service to their institution and profession.

In this Appendix, our goal is to compare how the importance of work activities for the management consultants under study relates to other knowledge work occupations.

We draw occupation data from the O*Net data set, a comprehensive database of occupational information from the US Bureau of Labor Statistics. O*Net includes information on skills, abilities, knowledge, work activities, and interests associated with over 900 occupations. The dataset has been extensively used in related studies (e.g., Felten et al. (2023); Eloundou et al. (2024)).

To accomplish this, we first identify the O*Net occupation corresponding to the management consultants in our study, the O*Net values that describe these consultants' work activities, and the O*Net occupations that qualify as knowledge work. Specifically, we aim to compare the concentration of the distributions of work activity importance across occupations.

F.1 O*Net Occupations

Our focal occupation is management consulting as practiced by large consulting firms (e.g., Boston Consulting Group, Deloitte, McKinsey & Company). The closest O*Net occupation is "`13-1111.00`", whose primary title is "Management Analysts." Occupation 13-1111.00 is described as:

> "Conduct organizational studies and evaluations, design systems and procedures, conduct work simplification and measurement studies, and prepare operations and procedures manuals to assist management in operating more efficiently and effectively. Includes program analysts and management consultants."

Occupation 13-1111.00 encompasses 39 occupation titles, which include 26 types of analyst, 7 types of consultant, 3 types of specialist, and 3 other titles. These are listed in table A.1.

F.2 O*Net Work Activity

The O*Net concept most closely related to our inquiry is "Generalized work activity," which is described as:

> "Work activities that are common across a very large number of occupations. They are performed in almost all job families and industries."

Title Root	Sub-type
Analyst	Administrative, Business, Business Development, Business Management, Business Operations, Business Process, Clerical Methods, Dealer, Employment Programs, Forms, Health Information Management Business, Health Information Management Data, Health Program, Health Systems, Human Resource, Industrial, Management, Management and Program, Operations, Performance Management, Program, Program Management, Project Management, Records Management, Reports, Survey
Consultant	Business , Business Management , Business Process , Organizational Development , Healthcare , Management , Performance
Specialist	Commercial, Health Program, Program Development
Management Scientist	
Management Systems Auditor	
Program Evaluator	

Table A.1: The 39 titles that are included under occupation code 13-1111.00.

There are 41 work activities associated with `13-1111.00` in the O*Net database.

Each work activity is associated with an "importance" and a "level." The importance metric uses a 5-point scale ("Not important," "Somewhat important," "Important," "Very important," "Extremely important"), while the level metric uses a 7-point scale ("1 Low," "2," "3," "4 Moderate," "5," "6," "7 High"). These values are collected sequentially: first, importance is evaluated. If an activity is rated above "Not Important," the level rating is then collected. Consequently, if an activity is rated "Not Important," the level is recorded as 0.

F.2.1 13-1111.00 Work Activities and Data Values

The table below lists some of these work activities for Management Analysts in descending order by their importance rating. The diversity of activities underscores how management analysts engage in a wide range of tasks that differ substantially from one another. In the analyses that follow, we focus on the importance metric, though our main conclusions remain the same if we instead use the level metric.

F.3 Knowledge Work

We are interested in comparing management consulting to other knowledge work occupations. To narrow our comparison set to knowledge work, we focus on occupations that require an education level at least as high as that of management consultants.

Element ID	Importance	Level	Description
4.A.1.a.1	4.70	5.48	Observing, receiving, and otherwise obtaining information from all relevant sources.
4.A.4.a.4	4.65	6.38	Developing constructive and cooperative working relationships with others, and maintaining them over time.
4.A.4.a.2	4.65	5.75	Providing information to supervisors, co-workers, and subordinates by telephone, in written form, e-mail, or in person.
4.A.2.b.1	4.62	6.00	Analyzing information and evaluating results to choose the best solution and solve problems.
4.A.4.b.6	4.45	6.38	Providing guidance and expert advice to management or other groups on technical, systems-, or process-related topics.
4.A.2.a.4	4.40	5.62	Identifying the underlying principles, reasons, or facts of information by breaking down information or data into separate parts.
4.A.2.b.4	4.38	5.43	Establishing long-range objectives and specifying the strategies and actions to achieve them.
4.A.4.a.1	4.38	4.71	Translating or explaining what information means and how it can be used.

Figure A.1: O*Net work activities and data values for consultants.

Education level serves as a practical proxy for knowledge work because higher-degree requirements reliably signal the need for advanced cognitive skills, specialized training, and the ability to handle complex or abstract problem-solving tasks. Occupations requiring at least a Bachelor's degree often entail analytical thinking, expert judgment, and creativity, all of which are hallmarks of knowledge work. Moreover, large-scale labor datasets—such as those from O*Net—commonly classify professional roles by their typical education prerequisites, making it straightforward to identify and compare these roles in an empirical, standardized manner.

O*Net collects information on required education levels under element "2.D.1," "Required Level Of Education (Categories 1-12)." The survey item asks respondents to select from 12 levels:

1. "Less than a High School Diploma"
2. "High School Diploma - or the equivalent (for example, GED)"
3. "Post-Secondary Certificate - awarded for training completed after high school (for example, in agriculture or natural resources, computer services, personal or culinary services, engineering technologies, healthcare, construction trades, mechanic and repair technologies, or precision production)"
4. "Some College Courses"
5. "Associate's Degree (or other 2-year degree)"
6. "Bachelor's Degree"
7. "Post-Baccalaureate Certificate - awarded for completion of an organized program of study; designed for people who have completed a Baccalaureate degree but do not meet the requirements of academic degrees carrying the title of Master."

8. "Master's Degree"
9. "Post-Master's Certificate - awarded for completion of an organized program of study; designed for people who have completed a Master's degree but do not meet the requirements of academic degrees at the doctoral level."
10. "First Professional Degree - awarded for completion of a program that: requires at least 2 years of college work before entrance into the program, includes a total of at least 6 academic years of work to complete, and provides all remaining academic requirements to begin practice in a profession."
11. "Doctoral Degree"
12. "Post-Doctoral Training"

These categories are coded from 1 to 12 in the database, with O*Net reporting the percentage of respondents who selected each level.

To compare education levels, we summarize each occupation's responses into a single value. For most occupations, the centrality measures differ only minimally. Thus, we use the median required education level. The centrality measures for consulting appear in table A.2.

Table A.2: **Management Consulting Education Level**

Occupation	Mean	Median	Mode
13-1111.00	6.38	6	6

Notes: The table presents a summary of required education level responses for consulting. Note that level 6 is defined as "Bachelor's Degree."

F.4 Data

The O*NET 29.1 Database contains 1016 occupation codes. Work activities are documented for 879 of these codes, and required education level data is available for 858 codes. Consequently, complete data is available for 858 occupations.

Since we aim to compare management consulting to other knowledge work occupations, we restrict our set to those requiring a median education level of at least 6 (i.e., "Bachelor's Degree"). There are 335 such occupations, which constitute our analysis data.

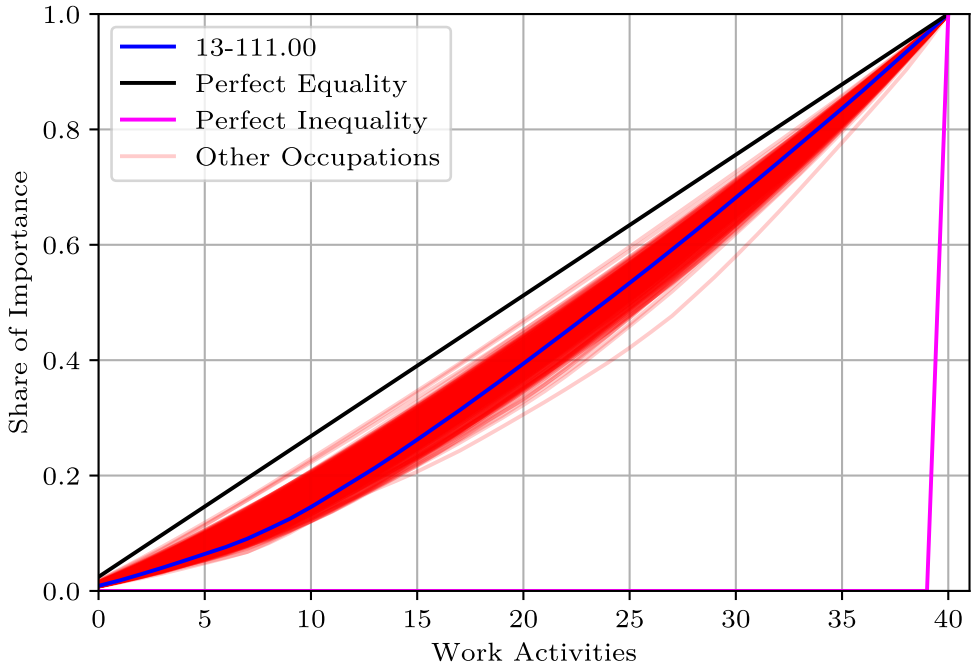
F.5 Concentration of Importance Distributions

To measure how importance is distributed across work activities, we use the Gini coefficient. This coefficient ranges from 0 to 1, where 0 indicates a perfectly even distribution (each activity is equally important) and 1 indicates a perfectly concentrated

distribution (only one activity holds importance). For occupation 13-1111.00, the Gini coefficient is 0.17.

Figure A.2 plots Lorentz curves for each of the 335 occupations in our analysis data. The total importance for each occupation is normalized to 1.0, and each curve connects 0 (the origin) to 1.0 (the 41st activity). Graphically, the Gini coefficient for an occupation is the ratio of the area between the 45-degree line and its Lorentz curve to the total area under the 45-degree line.

Figure A.2: **Lorentz Curves of the 335 Knowledge Work Occupations**

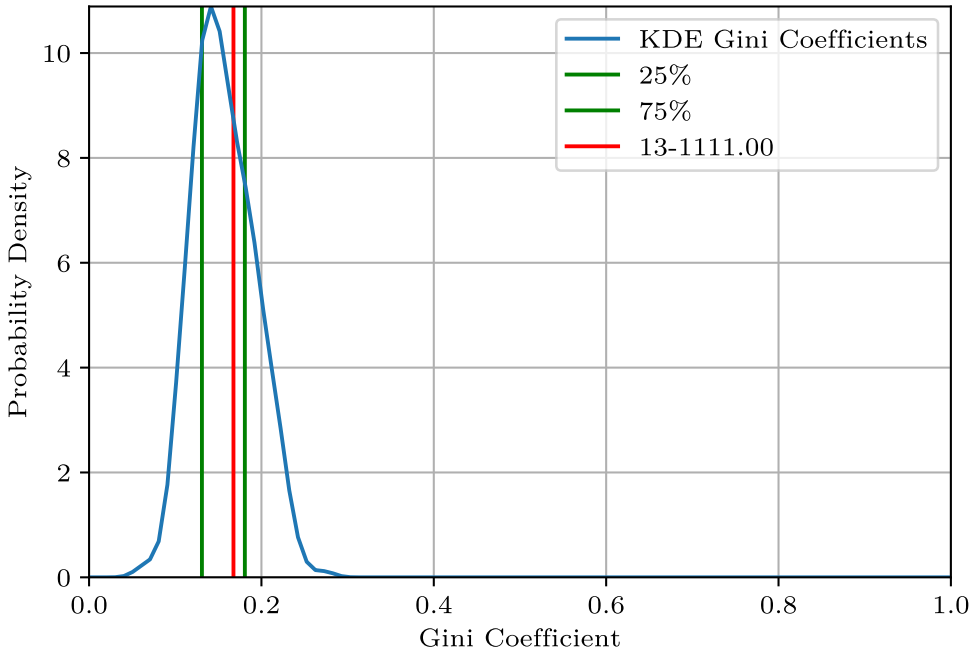


Notes: Each curve represents a Lorentz plot for one of the 335 occupations, illustrating how "importance" is distributed across work activities. A perfectly even distribution yields a straight diagonal line, while greater curvature indicates a more concentrated distribution of importance.

F.6 Importance Gini Coefficients Across Occupations

Figure A.3 shows the distribution of Gini coefficients among occupations whose education requirements are at least as rigorous as `13-1111.00` (n = 335). A kernel density estimate of the Gini coefficient distribution is provided, with the interquartile range and the Gini coefficient value for consulting highlighted. The concentration of work activity importance for consulting is unremarkable, falling within the interquartile range of Gini coefficients.

Figure A.3: **Gini Coefficients of Work Activity Importance**



Notes: This figure shows the distribution of Gini coefficients measuring how work-activity importance is distributed among occupations that require an education level at least as high as 13-1111.00 (n = 335). Lower Gini values indicate a more even distribution across activities, while higher values suggest a greater concentration of importance in fewer tasks.

Overall, this analysis shows that management consultants perform a wide range of activities that by their nature tend to be quite different from each other and difficult to compare. From this perspective, this is similar to what happens to knowledge work occupations more broadly. In fact, according to various measures of distribution of work activities, management consultants are not special compared to other knowledge work: we observe no clear evidence that consultants are special in terms of the variety or distribution of work activities they engage in. Rather, across relevant O*Net occupations that require a similar level of educational attainment, management consulting falls well within the normal range when measuring the importance of various tasks.

Although BCG consultants may be on the high end among their peers, from a broader labor market perspective, the job activities of 'management analysts' resemble those of

other occupations that require college-level education and specialized cognitive skills. This supports the potential generalization of our findings to knowledge work more broadly.

G Surveys with BCG Managing Directors and Partners

To further establish the practical applicability and real-world relevance of our experimental tasks, we conducted a survey with 11 Managing Directors and Partners (MDPs) at BCG. MDPs are among the most senior executives at BCG and serve as the primary evaluators of consultant performance. Their feedback offers three key insights into the robustness of our study design:

We find that the dimensions assessed in our inside-the-frontier and outside-the-frontier tasks are indeed pivotal for success at BCG, both during recruitment (figure A.4) and throughout a consultant's career (figure A.5). Importantly, these skills are not merely important for the interview process but directly inform consultants' daily work and long-term progression within the firm. These perspectives from senior leaders across BCG's global operations confirm that our experimental tasks closely mirror the real-world demands of management consulting.

These survey results strengthen the external validity of our findings. First, the MDPs perspectives represent a wide range of industry verticals and functional specialties, underscoring the broad applicability of our task design. Second, the MDPs confirmed that the tasks in our experiment capture real-world demands. Finally, because MDPs are the primary arbiters of promotion and advancement within the firm, their validation affirms that the qualities we tested bear directly on career trajectories.

Figure A.4: MDPs' views on the importance of each skill dimension for a successful application

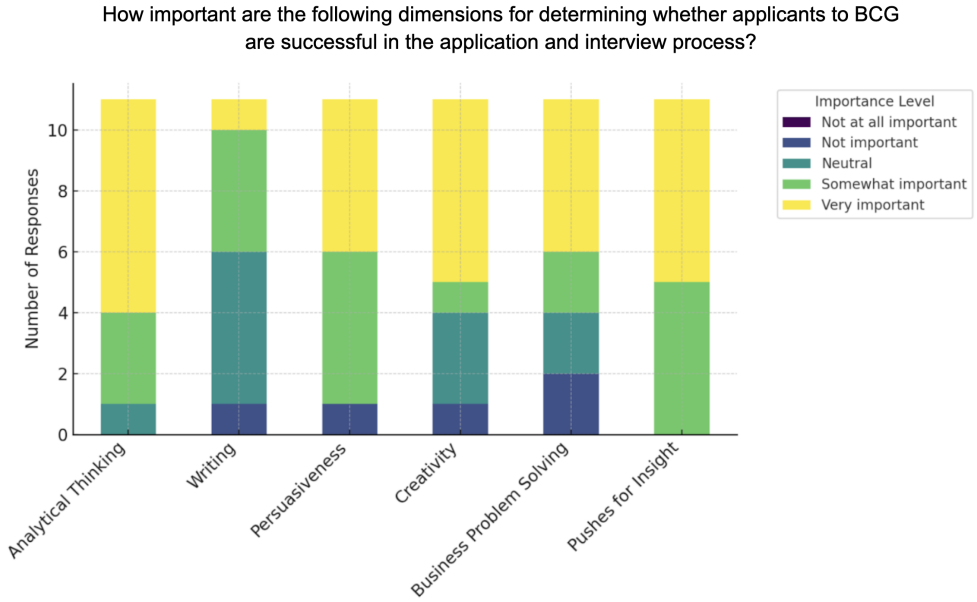
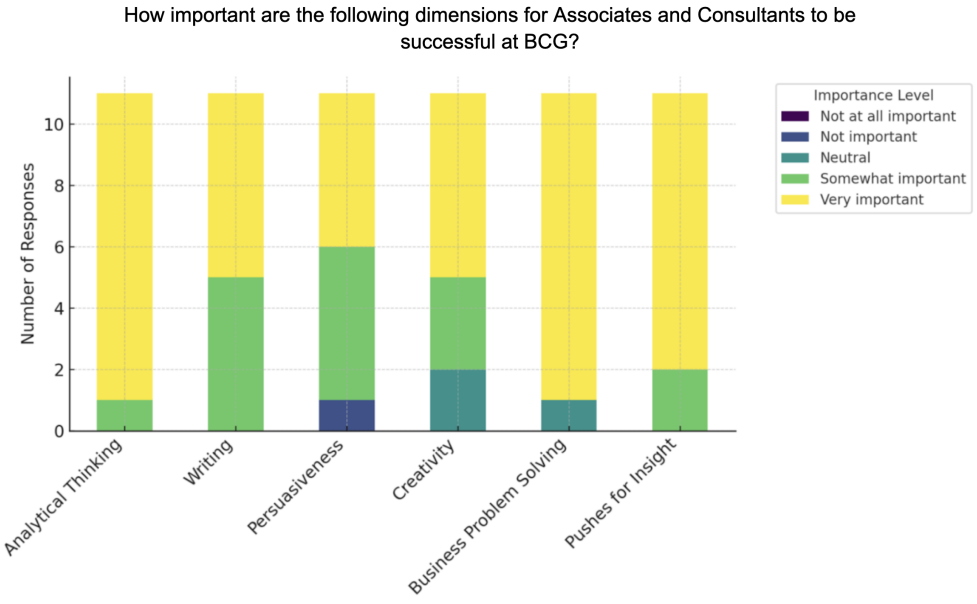
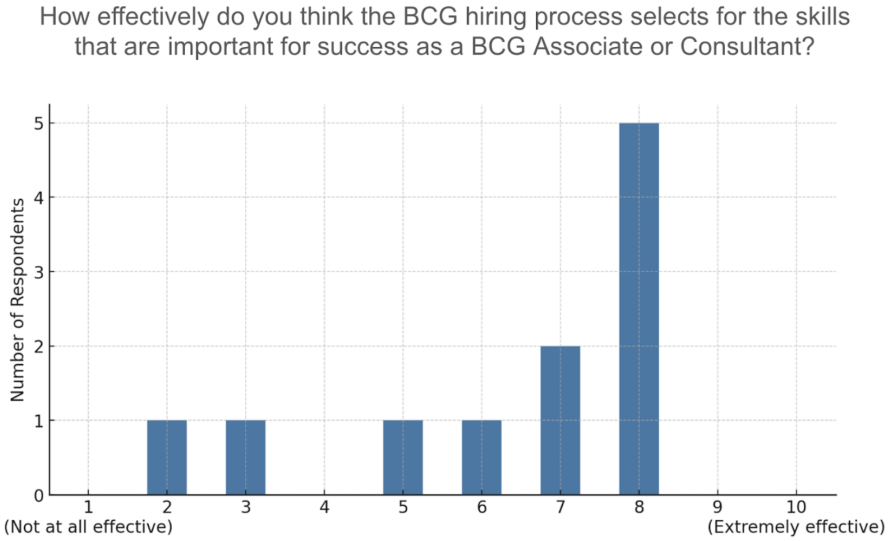


Figure A.5: MDPs' views on the importance of each skill dimension for on-the-job success



Finally, these senior executives strongly affirmed that BCG's interview and evaluation processes are effective at selecting for skills that consultants require to thrive in their professional roles. Figure A.6 reports the answers to a question we asked MDPs about the effectiveness of BCG's hiring. The median response is 7/10, demonstrating high satisfaction with the hiring process for selecting relevant consultant skills.

Figure A.6: MDPs' views on the effectiveness of the hiring process



H Robustness Check with Controls for Grammar and Spelling

Differences in perceived 'quality' of task completion may be explained by variations in spelling and grammar quality between groups rather than by deeper skills like creativity and persuasion. To address this concern, we constructed a control variable measuring the number and rate of spelling and grammar errors in responses.

Our analysis shows that while these errors do have a statistically significant effect on the outcome score, their effect size is small. As such, the treatment effect largely remains unchanged. We detail our analysis below.

H.1 Methods

In this analysis, we disaggregate the data for each response answered by the participant. The number of grammar errors is calculated separately for each response associated with a participant.

To detect errors in grammar and spelling, we leveraged a Python wrapper for LanguageTool, an open-source grammar tool known as the spellchecker for OpenOffice. LanguageTool detects a wide range of issues, including unpaired quotes, improper capitalization, subject-verb agreement errors, and word redundancies. We construct a control variable that, for each answer, counts the number of flags raised by the LanguageTool when parsing the answer. Below are two examples of flags that the LanguageTool raised:

Grammar/Spell Check Example
Context: "...dead lifting shoes,marathon runners etc."
Flag: whitespace
Message: Put a space after the comma
Replacements: [', marathon']

Grammar/Spell Check Example
Context: "Create shoes with unique functionality - like self-tying shoe laces..."
Flag: misspelling
Message: This word is normally spelled as one
Replacements: ['shoelaces']

The tool exhibits a high Type I error rate for proper nouns, which are prevalent in our dataset because participants frequently invent names for their shoe design products.

Grammar/Spell Check Example
Context: "FlexFit Racer: A lightweight, flexible shoe that adapts..."
Flag: misspelling
Message: Possible spelling mistake found
Replacements: ['Flex Fit']

For this reason, we filter out all misspelling flags when the first letter of the word is capitalized.

Below is a table showing the distribution of errors flagged within each group of the original dataset:

# Errors	0	1	2	3	4	5	6+
Control	49.6%	22.6%	12.8%	6.6%	2.8%	1.6%	4.0%
GPT + Overview	73.7%	16.7%	4.5%	2.0%	0.9%	0.9%	1.5%
GPT Only	72.8%	16.0%	5.8%	2.3%	1.2%	0.6%	1.3%

Using this data, we constructed control variables for the number of errors in a given answer, as well as the error rate as a portion of the number of words in the answer. The error rate is multiplied by 100 for easier interpretability (the coefficient represents the effect of a 1% increase in the grammar error rate). Standard errors are clustered by participant ID and question number.

H.2 Results

The robustness check on the original dataset is shown in the table below. Controls for the number of grammar errors, as well as the grammar error rate, do not substantially affect the coefficients or standard errors associated with either experimental condition. Thus, we can safely conclude that superficial grammar and spelling errors are not responsible for the treatment effect observed in the paper.

Table A.3: **Quality Robustness Check with Grammar Controls**

	(1) Quality	(2) Quality	(3) Quality
GPT + Overview	1.514***	1.589***	1.444***
	(0.140)	(0.133)	(0.143)
GPT Only	1.360***	1.435***	1.294***
	(0.140)	(0.133)	(0.143)
# Grammar Errors		0.115***	
		(0.025)	
Error Rate			-0.052***
			(0.014)
R2	0.241	0.259	0.252
Control = GPT + Overview	0.015	0.017	0.016
Control Mean	4.412	4.412	4.412
Observations	6127	6127	6127

I Interrater reliability - main grades

A key component of our study involves the evaluation of complex, knowledge-intensive work. For inside-the-frontier tasks, quality was assessed through subjective rubrics. Given the inherent subjectivity of such evaluations, it is crucial to establish the consistency and reliability of our graders. This appendix details the comprehensive inter-rater reliability (IRR) analyses we conducted to ensure the robustness and validity of our measurement approach.

I.1 ICC and Kendall's Tau

We began by calculating the Intraclass Correlation Coefficient (ICC) and Kendall's Tau. For individual footwear questions, we observed an average ICC(2,1) of 0.34. This reliability varied considerably across questions, ranging from a maximum of 0.59 to a minimum of 0.14. Additionally, Kendall's tau, another measure of consistency, averaged 0.38 across all grader pairs.[31]

I.2 Gwet's Agreement Coefficient (AC2)

To further probe the robustness of our graders' consistency, we recalculated intercoder reliability using Gwet's AC2. We chose AC2 (and not AC1) because it allows weighted agreement on ordinal scales. To reflect the point that substantial divergences are more consequential than small discrepancies, we apply quadratic weights. The resulting coefficient is 0.8066 (95 % CI = 0.8015–0.8117), reported below, a level conventionally interpreted as "almost perfect" agreement.

As a falsification check, rerunning the procedure on a dataset in which all grades were randomly permuted drives the coefficient to 0.004, confirming that the statistic behaves as theory predicts.

We next replicate the reliability checks on the three independent product-design experts who graded all 18 inside-frontier questions, again using quadratic weights. Their agreement too is strong (AC2=0.608,95% CI 0.594–0.623).

The apparent discrepancy with our previously reported average ICC (ICC(2,1) = 0.34) is explained by the specific properties of these two statistics. The ICC(2,1) model measures absolute agreement, so any systematic shift in how raters use the scale directly lowers the coefficient. This is because systematic variance between raters—such as one being consistently harsher than another—is implicitly included in the ICC's error term but assumed away. The coefficient thus penalizes differences in rater means and standard deviations, as opposed to only inconsistencies in their rankings. Such differences are present here, as we discuss below: while all raters used the full scale, their mean scores ranged from 4.0 to 7.1. Gwet's AC2 is robust to this issue because it computes expected chance-agreement using the actual marginal distributions of each rater, rather than assuming all rater marginals are the same (as does ICC) or that they are fully independent (as does Cohen's kappa), thereby decoupling agreement from systematic level differences

[31] Per question data is available upon request.

Figure A.7: **Gwet's AC2 - business graders**

Metric	Value
Coefficient Value	0.8066
Coefficient Name	AC2
95% Confidence Interval	0.8015, 0.8117
p_value	0
Z-Score	308.93
Standard Error (SE)	0.00261
Observed Agreement (Pa)	0.9552
Expected Agreement (Pe)	0.7685

Figure A.8: **Gwet's AC2 - shoe design experts**

Metric	Value
Coefficient Value	0.60852
Coefficient Name	AC2
95% Confidence Interval	0.59402-0.62301
p_value	0
Z-Score	82.30059
Standard Error (SE)	0.00739
Observed Agreement (Pa)	0.90684
Expected Agreement (Pe)	0.76204

across raters. We are thankful that you suggested this point, as we believe that we now have a more appropriate estimate for our setting.

I.3 Dispersion Among Coders

To provide a clearer picture of grader behavior, we analyzed the distribution and dispersion of scores. Figure A.9 below shows the overall distribution of scores across all business graders.[32] The histogram is close to normal, centered on 5–6, and exhibits ample spread—evidence that graders made full use of the 1-to-10 rubric rather than clustering at the extremes.

Figure A.9: **Distribution of Scores - Business Graders**

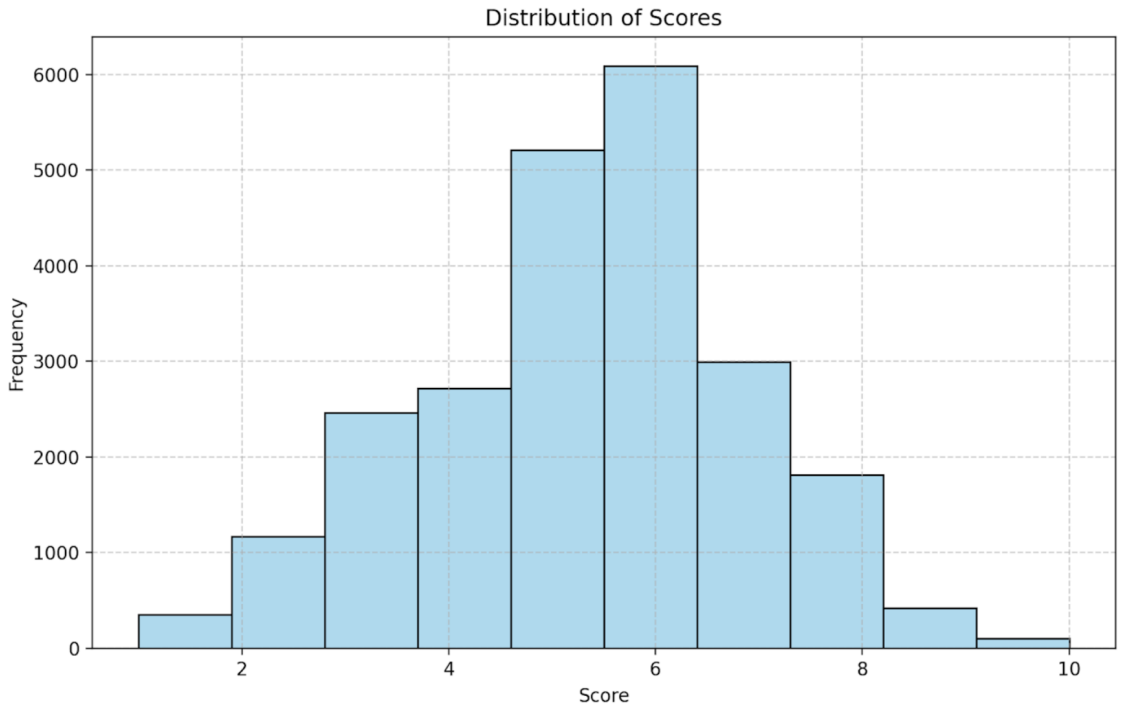


Figure A.10 overlays kernel-density curves for each of the ten graders. While individual raters differ somewhat in central tendency and variance, the curves reveal broad overlap and no outliers who rely on an idiosyncratic slice of the scale. This pattern reinforces the high Gwet's AC2 estimates reported above: disagreements tend to be small shifts around the shared modal range.

Figure A.11 summarizes statistics for each individual rater, again showing that while raters differed in their mean scoring tendencies (ranging from 4.0 to 7.1), they all utilized substantial portions of the 10-point scale with similar ranges.

[32] A very small number of graders mistakenly assigned scores of 0 instead of 1 ((approximately 0.7% of all grades). These instances were retained in the calculation of average grades used in the main analyses, but are excluded from the grader-level distributions. Rerunning the main analyses with these 0s excluded yields results that are fully consistent with those reported in the paper.

Figure A.10: **Distribution of Scores among Evaluators - Business Graders**

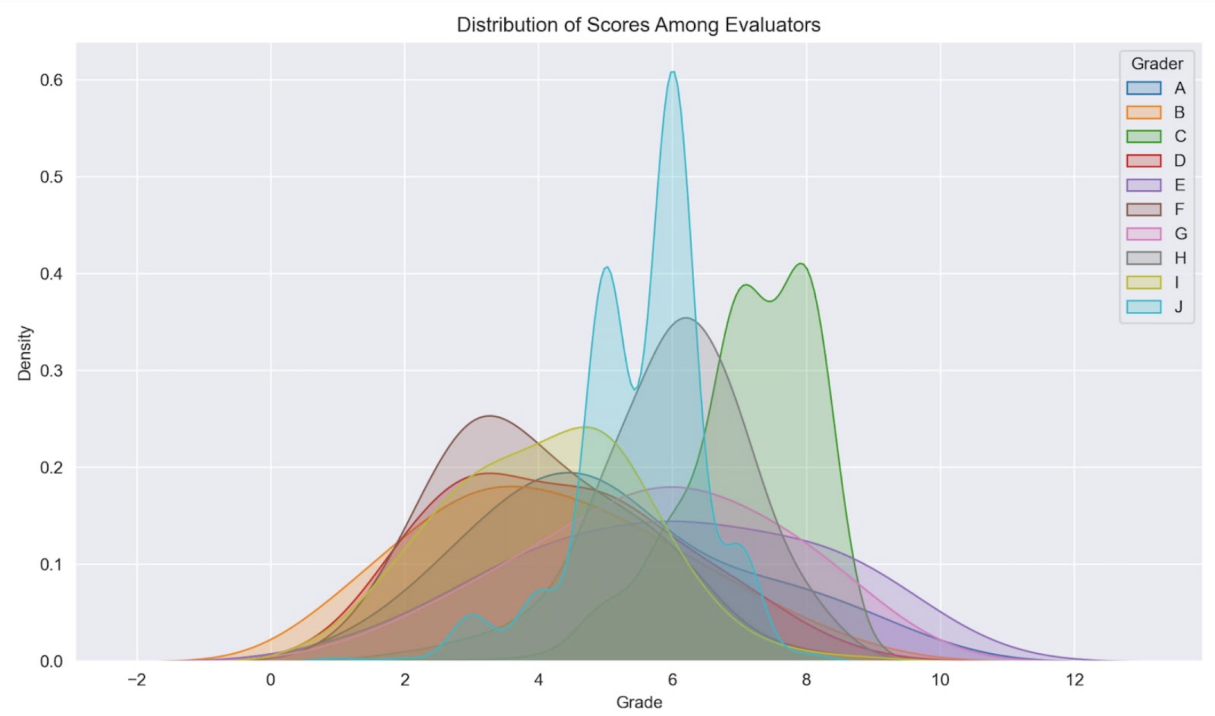


Figure A.11: **Summary Statistics - Business Graders**

Rater	N Ratings	Mean Score	Std Dev	Min Score	Max Score
A	3153	5.1	2.0	1	10
B	1387	4.2	1.9	1	9
C	2256	7.1	1.0	1	9
D	2453	4.3	1.7	1	10
E	1348	6.0	2.2	1	10
F	881	4.0	1.4	1	9
G	881	5.7	1.9	1	10
H	1373	6.0	1.1	1	9
I	1831	4.1	1.5	1	9
J	7781	5.5	1.0	1	9

Finally, figures A.12 and A.13 summarize the distribution of scores among the shoe design experts. As with the business raters, we observe systematic differences across graders in both central tendency and spread, with one evaluator consistently assigning higher marks. At the same time, all three use the full scale and exhibit substantial overlap in their scoring distributions, indicating broadly comparable grading practices.

Figure A.12: **Distribution of Scores among Evaluators - Business Graders**

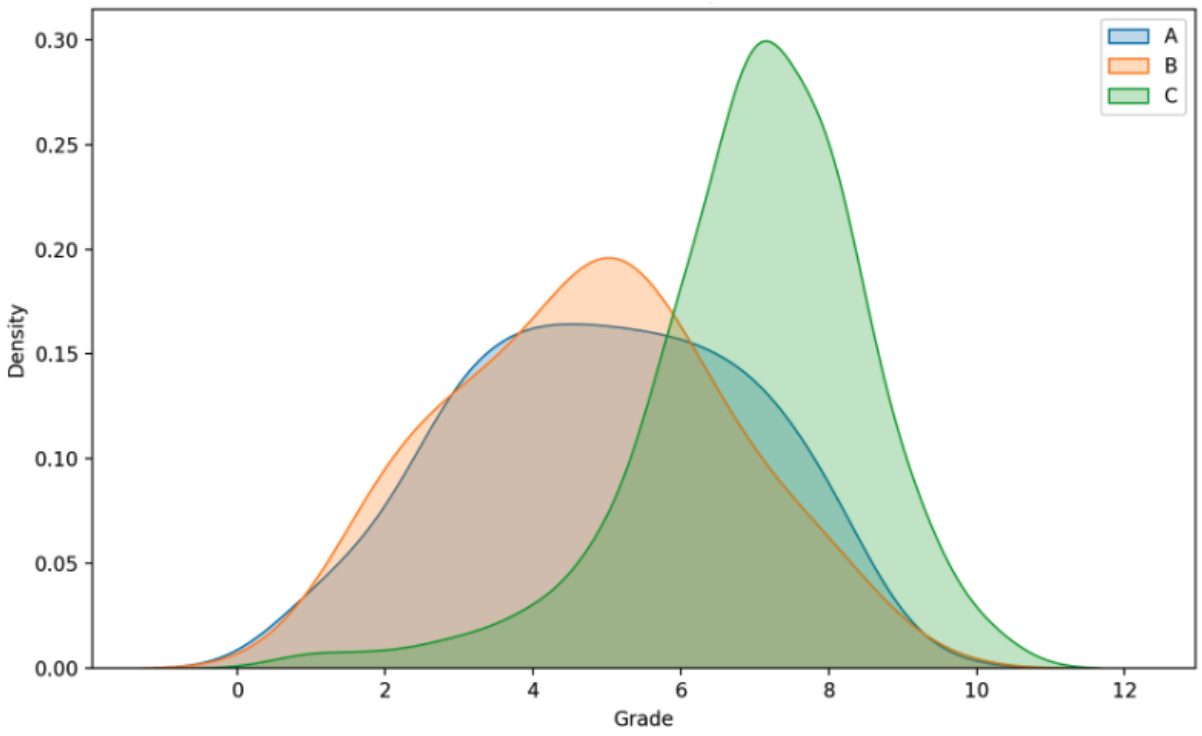


Figure A.13: **Summary Statistics - Business Graders**

Rater	N Ratings	Mean Score	Std Dev	Min Score	Max Score
A	5051	5	1.9	1	10
B	2728	4.8	1.9	1	10
C	4481	7	1.5	1	10

Taken together, these comprehensive analyses demonstrate a high degree of reliability and consistency in our evaluation data, strengthening the validity of our main findings.