Reflections from the Field

Comparative evaluation of an AI-powered life coach against traditional coaching methods

David Brown (Share Ventures, Los Angeles, USA)
Marlene Orozco (Stanford University, California, USA)
Noah Lloyd (Share Ventures, Los Angeles, USA)

Abstract

This study assesses the efficacy of “1440”, an AI-powered life coaching tool, in comparison to traditional human coaching. Utilizing a controlled experimental design, participants were divided into three distinct groups and engaged in a standardized coaching scenario. The study evaluates the performance of 1440 across multiple key metrics, including goal achievement, satisfaction, and perceived support. The empirical findings indicate that 1440 significantly outperforms traditional human coaching in several critical dimensions, suggesting its potential as a scalable and accessible alternative for personal development and professional growth.

Keywords

digital coaching, artificial intelligence, life coaching, personal development, coaching efficacy

Article history

Accepted for publication: 03 January 2025
Published online: 03 February 2025

Introduction

The integration of Artificial Intelligence (AI) in the domain of personal development and coaching introduces innovative pathways for enhancing the accessibility and efficiency of coaching services. “1440", an AI-driven coaching tool developed by Share Ventures, aims to replicate and potentially surpass the effectiveness of human coaching by providing continuous, contextual, and connected support. This research seeks to empirically evaluate the capabilities of "1440" in comparison to conventional coaching methods, focusing on key performance indicators pertinent to the coaching process.

As the coaching industry evolves, the demand for scalable, cost-effective, and accessible solutions has markedly increased. AI-powered coaching tools like “1440” present an opportunity to meet this demand by leveraging advanced technologies to deliver personalized coaching experiences. This study aims to contribute to the growing body of literature on AI in coaching by conducting a rigorous comparative analysis between AI-powered and traditional human coaching methods. By doing so, we seek to understand the potential benefits and limitations of AI in this context and to provide insights into the future of coaching practices.

Background

The concept of life coaching has evolved significantly, with a growing emphasis on accessibility and personalization. Traditionally, coaching has been a human-centric endeavor, relying on the coach's ability to build rapport, provide feedback, and facilitate the client's personal and professional growth. However, the advent of AI has introduced new possibilities for enhancing and scaling coaching services.

Evolution of Coaching and AI Integration

The coaching industry has seen various phases of evolution, from the early days of ad hoc, non-qualification-based training to the current era of evidence-based, professionalized coaching practices (Passmore & Woodward, 2023). This evolution has been driven by the need to standardize coaching practices and improve their effectiveness. AI-powered coaching represents the latest phase in this evolution, offering the potential to democratize access to coaching by making it more affordable and widely available (Terblanche, 2020).

AI in coaching is not an entirely new concept. Previous studies have explored AI's role in educational and therapeutic settings, with mixed outcomes. For instance, Terblanche and Cilliers (2020) identified that users' performance expectancy, social influence, and attitude significantly impact the adoption of AI coaches, although effort expectancy and perceived risk were less influential. Their research underscores the importance of developing AI coaching systems perceived as useful and socially endorsed.

Comparative Efficacy of AI and Human Coaching

Prior research on AI's efficacy in coaching has produced promising results. Terblanche et al. (2022) demonstrated that AI coaches could effectively improve goal attainment, similar to human coaches, but highlighted the need for integrating supportive elements such as empathy, which AI lacks. Mai et al. (2022) discussed the impact of a chatbot's disclosure behavior on the working alliance and acceptance, suggesting that AI's human-like interaction capabilities significantly influence user engagement.

The Coaching Cube framework by Segers et al. (2011) offers a comprehensive model to understand various coaching dimensions, emphasizing the importance of matching coaching agendas, coach characteristics, and approaches to enhance effectiveness. The integration of multimedia, flexibility in interaction methods, and the ability to personalize coaching interactions are critical factors that can enhance the effectiveness of AI coaching tools.

Ethical Considerations in AI Coaching

Ethical considerations are paramount in the development and deployment of AI coaching tools. Hannafey and Vitulano (2013) highlight the importance of addressing issues such as data security, transparency, and conflicts of interest. Ensuring the ethical use of AI in coaching involves maintaining confidentiality, providing clear communication about data usage, and implementing safeguards to prevent bias and misuse of information.

Ethical AI coaching requires transparency in how AI systems are trained and used, as well as adherence to professional standards that protect the client's interests. The development of AI coaches should involve continuous monitoring and evaluation to ensure they meet ethical guidelines and provide a safe, effective coaching experience.

Methodology

Participants

One hundred eighty participants were recruited from UserTesting.com and randomly assigned to one of six groups across twelve sessions. The six groups were each formed of two fifteen person sessions (30 total). The participants were young adults, diverse in gender, and professional background to ensure the generalizability of the findings. All participants were Americans over the age of 18 years old with a max age of 31 and an average of 25.8-years-old. Annual household income (interpreted from ranges) is approximately $75,944. Participants were compensated $10 for their participation in the study. One participant was removed from the “video group” for not fully completing the experiment.

Design

Each participant from UserTesting signed up for the study and voluntarily completed the experiment in return for compensation. UserTesting is a platform that enables people to complete surveys and experiments in exchange for money. It is standard for compensation of the experiment to scale with the duration of the study. Candidates on the UserTesting platform are generally doing so to earn either a primary or additional source of income within the gig-economy. The study employed the following six groups. Each participant was presented with one of these samples.

1. Video Group

Viewed the actual coaching session between an ICF master qualified coach and their client (shared via a link to Vimeo). The coach and client were aware the session was being recorded and publicly shared. The interaction was approximately 26 minutes in length (with options to watch at an increased speed), and covered culturally competent topics related to family finances, children, upbringing, and navigating marital relationships. This coaching session was also relevant to a time sensitive binary decision to be made by the client.

2. Human Transcript

Read a transcript of the same video, processed through GPT 4 to remove "ums", "ahs", and dead-end sentences in order to bring it more in line with the written transcript of group 3.

3. 1440 Transcript

A transcript was shown to study participants where the coach was replaced by 1440’s coach Nia (an AI built atop a proprietary LLM framework) and the client was role played by one of the staff members.

4. Unmodified GPT-4 Transcript Group

Participants are presented with a transcript of an employee roleplaying as the client from the video while the coach is represented by an unmodified GPT-4 instance.

5. GPT-4 “Coach Nia” Persona

Before the interaction began, a GPT native tool was used to inform the LLM how to respond, we fed it the core guidelines for how our own coach Nia responds. However, this GPT instance with a Nia persona, did not have the underlying frameworks or tech stacks that Nia has within the 1440 platform. After this setup, the conversation proceeded as with prior examples: the same employee role played the client scenario and study participants were presented with the transcript.

6. GPT-4 “Meta” Coach

In this version of the GPT-4 experiment, the employee playing the client persona advised GPT on how best to respond to them while they were being coached. For example, the client may send a message, and upon receiving say “write to me more succinctly” and continue the conversation upon receiving a response in line with their expectation.

Measures and Procedures

Each of the five of the six groups were asked a total of nineteen questions. The group presented with the real video was presented with seventeen questions and skipped the final two, which asked questions about if the coach in the study was a human or AI due to the video making clear the coach was human. The first six questions were presented before being presented with the coaching scenario and were related to general thoughts and preconceived notions about coaching in addition to clarifying the definition of a coach in this context, which we referred to as non-athletic, non-activity-specific advising: “You may have heard of them, or their contemporaries referred to as a life coach, executive coach, counselor, or therapist.” Users were asked if they had received coaching before, if they know anyone who has received coaching, the percentage of the population they believe would benefit from a coach, the age they believe coaching should start, and pre-existing concerns they have with coaches.

The groups were then presented with their corresponding transcript via a view only Google Doc except for the “Video” group who was presented with a link to the video on Vimeo. After being exposed to the coaching sample, the participants were asked to evaluate the coach across eight variables: empathy, effectiveness, competence, communication, approachability, problem-solving, overall quality, and value to client. Each variable is shown to participants with a brief longer expounding, portrayed below, to help specify the measure requested of the users. When presented to participants, they were given longer titles. Variables were selected based on their relevance to the coaching process and their ability to provide a holistic assessment of the coaching experience. “The eight variables were evaluated on (due to technical constraints) on a scale from 1 - 11, and later corrected to a scale of 0 - 10. Value to client" was later removed from data analysis because without deeper product information around price and accessibility (which could not be introduced without de-blinding the study), the metric was considered too close to “overall quality”.

Empathy: Coach’s empathy and ability to understand client concerns

Empathy is a fundamental component of effective coaching, as it allows coaches to understand and respond to the emotional states of their clients, fostering a supportive and trusting relationship (Ianiro, Lehmann-Willenbrock, & Kauffeld, 2015). This aspect is supported by further studies emphasizing the role of empathy in establishing a strong coach-client bond, which is crucial for the coaching process (De Haan, Sills, & Knight, 2016).

Effectiveness: Effectiveness in guiding the client towards their goals

Effectiveness assesses the overall impact of coaching on achieving desired outcomes, reflecting the coach's ability to facilitate meaningful change (Terblanche et al., 2022). Research by Grant (2014) highlights the importance of goal attainment and client satisfaction as primary indicators of coaching effectiveness.

Competence: Display of competence and expertise

Competence relates to the coach's knowledge, skills, and abilities to provide high-quality coaching services, ensuring that the coach can address a wide range of client needs (Segers et al., 2011). Passmore (2010) notes that a competent coach should possess a robust understanding of coaching methodologies and exhibit continuous professional development.

Communication: Communication skills, particularly in clarity and adaptability to client needs

Communication is essential for successful coaching, as it involves clear, open, and effective exchanges of information between the coach and the client, which are vital for goal-setting and progress tracking (Passmore & Woodward, 2023). Effective communication has been shown to enhance the coaching relationship and facilitate better client outcomes (Boyce, Jackson, & Neal, 2010).

Approachability: Approachability and the ability to create a safe space for the client

Approachability refers to the coach's ability to create a welcoming and non-judgmental environment, encouraging clients to openly share their thoughts and concerns (Terblanche & Cilliers, 2020). A coach's approachability is linked to the creation of a safe space, which is essential for client engagement and trust (Baron & Morin, 2009).

Problem-solving: Problem-solving skills and the capacity to provide actionable advice

Problem-solving is a key aspect of coaching, as it involves the coach's ability to help clients identify issues and develop strategies to overcome them (Mai et al., 2022). Research by Theeboom, Beersma, and Van Vianen (2014) indicates that problem-solving skills are critical for facilitating clients' personal and professional development.

Overall Quality: Overall quality of the coaching session

Overall quality encompasses the general satisfaction with the coaching experience, integrating various aspects of the interaction into a single comprehensive measure (Terblanche, 2020). Evaluating overall quality helps in understanding the holistic impact of coaching and is supported by studies that assess client perceptions of coaching effectiveness (Jones, Woods, & Guillaume, 2016).

Data Analysis

Data was exported from UserTesting into spreadsheets. The data was collected and collated within Google Sheets. Significance analysis was completed through a series of one tailed t-tests which evaluated the non-adjusted numbers from each group. The significance threshold for this experiment was a p-value of < 0.05.

Results

1440 outperformed the apples-to-apples comparison to the human transcript of the coaching session in all seven of the categories with statistical significance (one tailed, T-Test p-value < 0.05) in all categories other than approachability. 1440 failed to reach a statistically significant improvement in approachability compared to any group.

The real coach generally performed better than its corresponding manuscript. The video particularly excelled in approachability where it scored higher than all other groups. 1440 was evaluated more highly than the video in six of the seven categories, though it only earned statistical significance in two of the seven. The video coach was evaluated to have performed significantly worse in measures of effectiveness and problem solving.

The unmodified instance of GPT 4, was outperformed by 1440 in five of seven categories, but only significantly so in two of the seven. 1440 performed significantly better in measures of effectiveness and communication. GPT 4 and 1440 performed the same on measures of approachability, and GPT 4 excelled in evaluations of competence - where it outperformed all other groups.

The version of GPT 4 that was given the input instructions from 1440 without its overarching infrastructure, the “Coach Nia Persona” underperformed 1440 in every category and significantly so in the same categories as the unmodified version of GPT 4: effectiveness and communication. Further, this persona of GPT-4 scored higher than its unmodified version in empathy but underperformed the unmodified version in all other categories.

The group that was given a transcript with a GPT-4 coach receiving meta level corrections evaluated the coach they were given as worse in every category when compared to 1440. However, 1440 was only significantly in two categories - empathy and overall quality. In comparison to the other GPT-4 groups, the Nia Persona group evaluated their coach as overall worse, with comparatively better performance in the areas where the other two GPT4 groups were significantly underperforming 1440.

Participants outside of the video scenario were also asked if they believed the coach they were presented with was a human or AI. 13% of participants incorrectly identified the human coach in the transcript as an AI. 1440’s transcript was predicted to be human by 16 of 30 group participants. Its Nia persona in GPT 4 scored the second best of the AIs with 33% believing that the AI coach was a human. The unmodified and meta version of GPT 4 scored lower with 10% and 7% respectively.

Figure 1: Comparative Evaluation of 1440 and a Master Certified Human Coach, statistical significance (0.05) outlined in black.

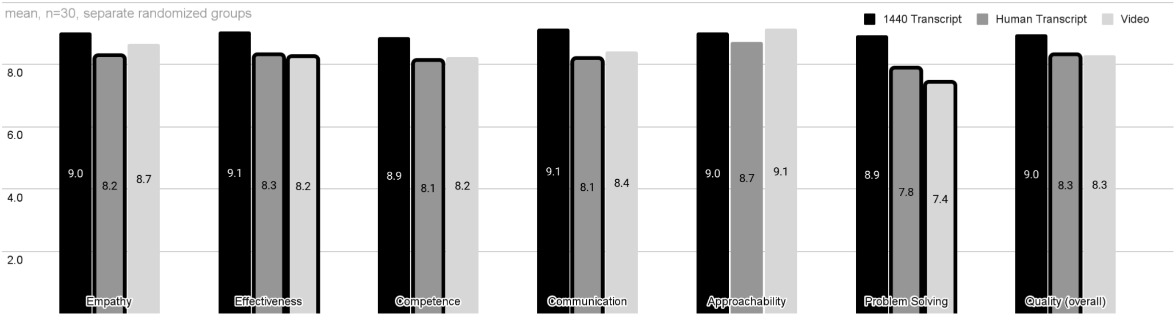


Figure 2: Comparative Evaluation of 1440 and GPT-4 Variants, statistical significance (0.05) outlined in black

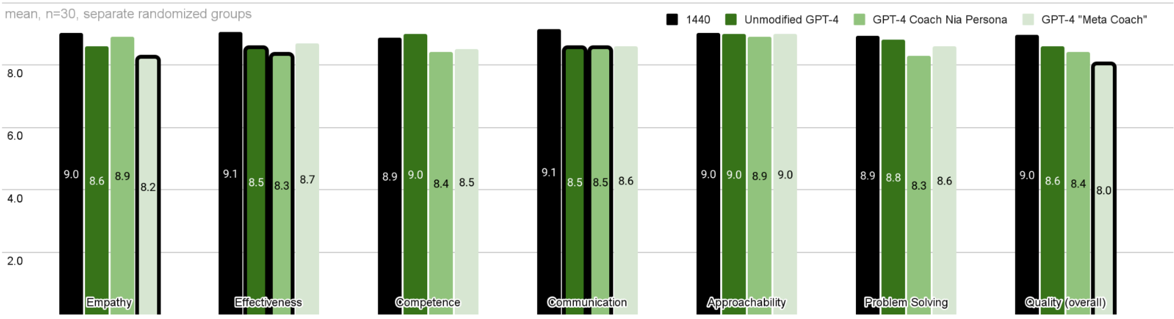


Figure 3: Percentage of Participants Who Identified the Coach in the Transcript as Human

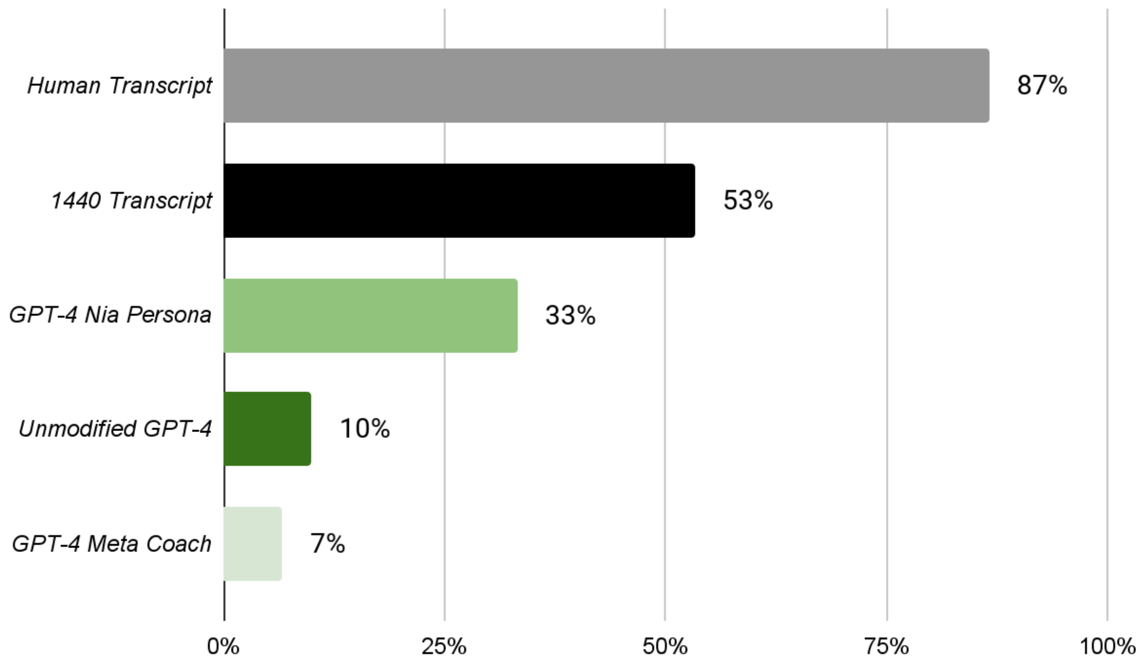
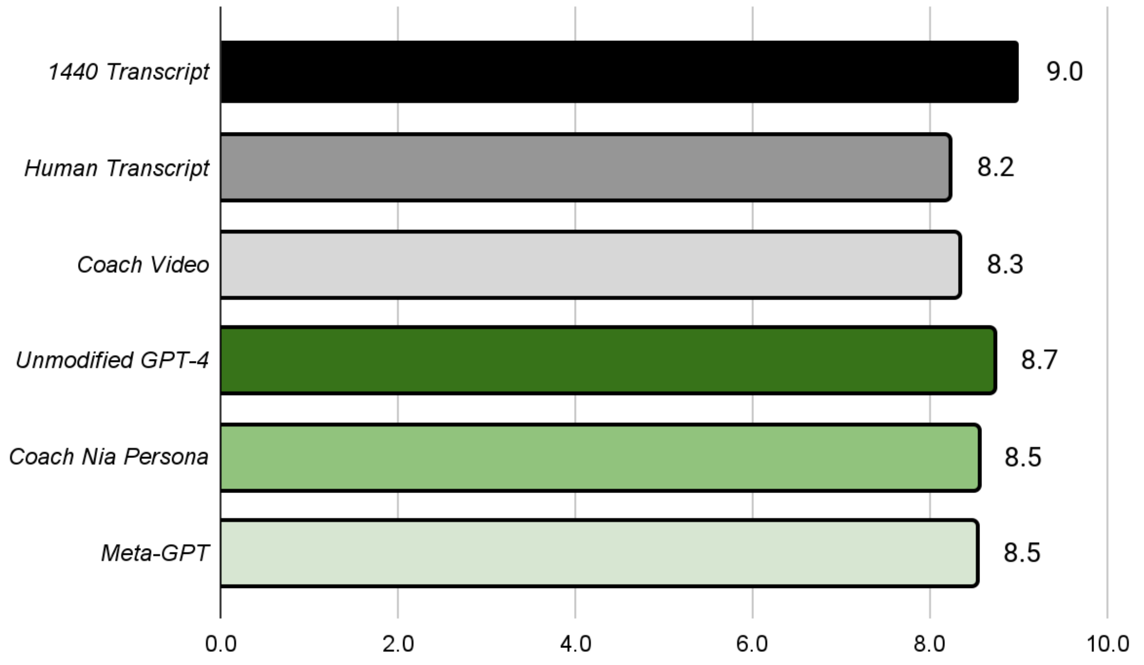


Figure 4: Mean Evaluation of All Categories, Equally Weighted



Discussion

The findings suggest serious potential for AI coaches as substitutes or additions to human coaching with the ability to match and even surpass experienced coaches in certain components of their work. 1440 was the highest overall performing group, surpassing both humans and Open AIs popular GPT-4. There were also several observable trends within the data when evaluating the AIs together in comparison to the humans, which may give us opportunities to read additional insights into the minds of the evaluators. The AIs perform better on measures of problem solving, competence, and communication. It is possible that these results form a more "transparent" and solution-oriented approach to coaching compared to the human in the study. Many coaches believe that their role as an advisor is to assess and guide through questioning, which in practice may lead to them holding back information or their full perspective during an interaction. LLMs may also be more likely to casually site sources or explain where and why it is making certain statements when compared to humans. Variances in communication styles are not necessarily differences in capabilities or knowledge, but rather a stance on transparency.

Both human coaches and 1440 have a case for this experiment not highlighting all of their greatest strengths. For this experiment an isolated training session was selected, which would remove the ability for the human coach to build a long-term relationship of mutual respect and understanding with the client. However, 1440 also has a variety of technological features and integrations which were not able to be integrated into this study. Another boon for 1440 and other LLM-based coaches is that they can have higher up-time and lower prices than humans are able to provide. 1440 aims to be available 24/7 whereas human coaches are normally spoken to for 1 - 4 interactions per month.

These results could have changed in unexpected ways were either the client, scenario, or coach replaced. People come to coaches with an incredibly broad array of problems (or lack thereof). Similarly, the diversity in coaches is almost as large as the clients that they see. Coaches vary in their areas of expertise, their backgrounds, training, certifications, methods, price points, and clientele. Further, not all great coaches are for every client. With any provider of mental health services, there is a component of "chemistry" or a provider's ability to intuitively understand the spoken and unconscious needs and experiences of the client. It is not possible for us to know what level of client/coach chemistry was present in the original video example. Similarly, we are not able to know from this data if the chosen scenario was either within the LLM's area of strength or weakness.

Limitations and Future Research

This study had several limiting factors including sample size, participant recruitment, the reliance on a single hand-picked scenario, and our team's ability to reliably role play on the client's behalf to the different LLMs present. Additionally, because neither LLMs nor humans consistently give identical output for identical input, there are inherent limits to replicability. While we aimed to recruit for a diverse set of academic, professional and financial backgrounds within our participants, gig-economy style services that offer payment in this manner does not have an equal value proposition to all adults.

Evaluating any coach is a difficult endeavor. True coaching interactions are generally built upon months or years of prior correspondence which is not something that we were able to present to users within this experiment. To that end, the research team elected for a video where it seemed the coach and client had no prior experience with one another, which impacts the professional's ability to do his job, and potentially the client's ability to be a successful client. Furthermore, there are difficulties around particular verbiage and expectations of what a coach should and should not be doing.1440 performed better on measures of problem solving than a human coach. It is also true that some schools of coaching believe that problems are to be solved by the client through adept questioning and not the coach. To that end, this school of coaching may believe that "problem solving" is an inappropriate measure to evaluate a coach's effectiveness from. Additionally, in choosing the audience, the team elected to pursue a crowd that more closely reflected the clients of coaching rather than a set of study participants that were peers or colleagues within the coaching space. This comes with an inherent trade-off of accuracy for precision.

We hope to conduct research in the future that accommodates longer term coaching interactions in environments that enables coaches to reflect the totality of their offerings. Additionally, alongside longer-term testing, we are hoping to do experiments where the frame of assessment is the impact on the client's life rather than the perceived quality of the coach.

Conclusion

1440 demonstrates significant promise as a viable alternative to traditional life coaching. In this controlled experiment it significantly outperformed every compared group in at least two of the seven measured variables, and it was the only AI where the majority of participants believed that it was a human. Additionally, coaching AIs like 1440 have the potential to offer scalable, lower cost, solutions to traditional coaching. This and other similar services could be used as another option of coaches to be selected from, similar to how any human coach would be selected. These services could also work in conjunction with human coaches, or due to their unique strengths have an opportunity to be scaled in places and communities with acute provider shortage as measure of or adjacent to public health.

Acknowledgements

The authors express gratitude to the study participants and UserTesting for facilitating the participant recruitment and data collection process.

References

Baron, L., & Morin, L. (2009). The coach-coachee relationship in executive coaching: A field study. *Human Resource Development Quarterly*, 20(1), 85-106. DOI: 10.1002/hrdq.20009.

Boyce, L. A., Jackson, R. J., & Neal, L. J. (2010). Building successful leadership coaching relationships: Examining impact of matching criteria in a leadership coaching program. *Journal of Management Development*, 29(10), 914-931. DOI: 10.1108/02621711011084231.

De Haan, E., Sills, C., & Knight, S. (2016). The coaching relationship: Putting people first. In S. Bachkirova, T. Bachkirova, G. Spence, & D. Drake (Eds.), *The Wiley-Blackwell Handbook of the Psychology of Coaching and Mentoring* (pp. 21-42). John Wiley & Sons. DOI: 10.1002/9781119237909.ch2.

Grant, A. M. (2014). The efficacy of executive coaching in times of organizational change. *Journal of Change Management*, 14(2), 258-280. DOI: 10.1080/14697017.2013.805159.

Ianiro, P. M., Lehmann-Willenbrock, N., & Kauffeld, S. (2015). Coaches and clients in action: A sequential analysis of interpersonal coach and client behavior. *Journal of Business and Psychology*, 30(3), 435-456. DOI: 10.1007/s10869-014-9374-5.

Jones, R. J., Woods, S. A., & Guillaume, Y. R. (2016). The effectiveness of workplace coaching: A meta-analysis of learning and performance outcomes from coaching. *Journal of Occupational and Organizational Psychology*, 89(2), 249-277. DOI: 10.1111/joop.12119.

Mai, V., Bauer, A., Deggelmann, C., et al. (2022). AI-based coaching: Impact of a chatbot's disclosure behavior on the working alliance and acceptance. In J. Y. C. Chen, G. Fragomeni, & A. Canossa (Eds.), *HCI International 2022—Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence* (pp. 391-406). Springer. DOI: 10.1007/978-3-031-21707-4_28.

Passmore, J. (2010). A grounded theory study of the coachee experience: The implications for training and practice in coaching psychology. *International Coaching Psychology Review*, 5(1), 48-62. DOI: 10.53841/bpsicpr.2010.5.1.48.

Passmore, J., & Woodward, W. (2023). Coaching education: Wake up to the new digital and AI coaching revolution! *International Coaching Psychology Review*, 18(1), 58-72. DOI: 10.53841/bpsicpr.2023.18.1.58.

Passmore, J., Diller, S. J., Isaacson, S., & Brantl, M. (Eds.). (2024). *The Digital and AI Coaches' Handbook: The Complete Guide to the Use of Online, AI, and Technology in Coaching*. Routledge. DOI: 10.4324/9781032469041.

Segers, J., Vloeberghs, D., Henderickx, E., & Inceoglu, I. (2011). Structuring and understanding the coaching industry: The coaching cube. *Academy of Management Learning & Education*, 10(2), 204-221. DOI: 10.5465/amle.10.2.zqr204.

Terblanche, N. (2020). A design framework to create artificial intelligence coaches. *International Journal of Evidence Based Coaching and Mentoring*, 18(2), 152-165. DOI: 10.24384/b7gs-3h05.

Terblanche, N., Molyn, J., De Haan, E., & Nilsson, V. O. (2022). Coaching at scale: Investigating the efficacy of artificial intelligence coaching. *International Journal of Evidence Based Coaching and Mentoring*, 20(2), 20-36. DOI: 10.24384/5cgf-ab69.

Terblanche, N., & Cilliers, D. (2020). Factors that influence users' adoption of being coached by an artificial intelligence coach. *Philosophy of Coaching*, 5(1), 61-70. DOI: 10.22316/poc/05.1.06.

Theeboom, T., Beersma, B., & Van Vianen, A. E. (2014). Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. *The Journal of Positive Psychology*, 9(1), 1-18. DOI: 10.1080/17439760.2013.837499.

About the authors

David Brown, Chief of Staff at Share Ventures received his bachelor's degree from Claremont McKenna in neuroscience and focuses his academic work in human behavior.

Dr. Marlene Orozco, Head of Research at Share Ventures, and Stanford GSB Research Fellow, specializes in reducing bias in entrepreneurship through advanced mixed methods research.

Noah Lloyd, Venture Manager at Share Ventures, is a full stack software engineer, data scientist, and licensed life coach.