Human Resource Development International

ISSN: 1367-8868 (Print) 1469-8374 (Online) Journal homepage: www.tandfonline.com/journals/rhrd20

A randomised controlled comparison of the effectiveness of human and AI chatbot coaching with goal attainment, wellbeing and self-efficacy

Erik de Haan, Nicky Terblanche & Kenneth Nowack

To cite this article: Erik de Haan, Nicky Terblanche & Kenneth Nowack (28 Feb 2026): A randomised controlled comparison of the effectiveness of human and AI chatbot coaching with goal attainment, wellbeing and self-efficacy, Human Resource Development International, DOI: 10.1080/13678868.2026.2633990

To link to this article: https://doi.org/10.1080/13678868.2026.2633990

HUMAN RESOURCE DEVELOPMENT INTERNATIONAL
https://doi.org/10.1080/13678868.2026.2633990

A randomised controlled comparison of the effectiveness of human and AI chatbot coaching with goal attainment, wellbeing and self-efficacy

Erik de Haan [a,b], Nicky Terblanche [c] and Kenneth Nowack [d]

[a]Ashridge Centre for Coaching, Hult International Business School, Berkhamsted, UK; [b]School of Business and Economics, VU University/Vrije Universiteit, Amsterdam, The Netherlands; [c]Stellenbosch Business School, South Africa; [d]Envisia Learning. Inc., Los Angeles, CA, USA

ABSTRACT

Executive coaching is widely used as a tool to develop leaders and organisations. The profession is rapidly changing due to the introduction and explosion of AI chatbots, AI-enabled co-pilots and digital technology platforms. This raises the question of how AI coaching compares to human coaching. In this randomised controlled experiment with $N = 114$ coachees, accredited human coaches were directly compared with automated AI coaches for the first time. Outcomes were measured using established and validated progress psychometrics on coachees' goals, motivation, resilience and wellbeing, and valid, momentary measurements. Our study emphasises a co-regulation model, which predicts better results for human coaches and minimal impact on attrition and dropout. The study demonstrated substantial effectiveness in diverse coaching outcomes but only for human coaching, with effect sizes in the mid-to-high range and strong predictions of success by initial self-efficacy and hope of the coachee. Implications and limitations of these findings for HRD theory and practice are discussed.

ARTICLE HISTORY
Received 5 May 2025
Accepted 23 January 2026

KEYWORDS
Randomised controlled trial; executive coaching; coaching outcomes; AI coaching; common factors; goal attainment

Introduction

Within a variety of popular HRD programmes, executive coaching is increasingly used as a tool to develop leaders and organisations and is estimated to be a $4.56 billion global industry (Ellinger & Kim, 2014; ICF, 2023). The market is transforming into larger, global providers of workplace coaching, online platforms and a variety of vendor-driven tools which attempt to consolidate, democratise and expand coaching services to more employees at all levels (Graßmann & Schermuly, 2021). In fact, the 2023 ICF Global Coaching survey reported that 81% of all respondents believe that digital coaching platforms can make coaching more accessible and provide a cost-effective alternative to more expensive live coaching programmes for high-potential and higher-level leaders.

Can AI digital coaching tools augment or effectively replace traditional face-to-face interactions (Bridgeman & Giraldez-Hayes, 2024; De Haan, 2024) in addressing specific presenting issues of employees in HRD leadership interventions? The answer might not

CONTACT Erik de Haan erik.dehaan@ashridge.org.uk

be entirely clear as this important question should not only be framed in terms of savings or accessibility enhancements but also in terms of which approach to coaching is best for specific individuals, with their unique presenting issues and the type of desired outcomes needed (e.g. goal achievement, skills development and long-term habit change).

The HRD field has embraced coaching as part of strategic human resource management, human capital and organisational development to enhance leadership capabilities and drive longer-term success. Successful coaching programmes are built on the HRD goal-performance models of Locke and Latham (2002, 2019) to ensure successful goal setting, deliberate practice back on the job and both meaningful personal and job-relevant work outcomes. Additionally, coaching, whether offered in person or via (AI) digital platforms, shares virtually the same intended purposes and processes as contemporary HRD (Hamlin et al., 2008).

With the maturing of the digital AI HRD era upon us (Ekuma, 2024; McLean & González Ortiz de Zárate, 2024; Passmore et al., 2025; Tavis & Woodward, 2024), it is important to determine whether attempts to democratise coaching within organisations will lead to successful goal achievement, habit and behaviour change outcomes in coachees whether in person or remote. As in other industries, AI coaching could offer specific advantages with respect to cost saving, democratisation, accessibility, innovation and a variety of interventions (e.g. Kanbach et al., 2024, with a case study around emerging clinical AI chatbots).

A wide variety of research into coaching processes and outcomes, such as goal achievement, employee well-being and performance, has now been undertaken (for recent meta-analyses, see Athanasopoulou & Dopson, 2018; Burt & Talati, 2017; De Haan & Nilsson, 2023; Jones et al., 2015; Nicolau et al., 2023; Theeboom et al., 2014). However, as Liu and Batt (2010) and Ianiro and De Haan (2024) have pointed out, much of the research into executive coaching has used psychology students or managers attending (MBA) courses as subjects, as opposed to experienced managers in a more realistic, ecologically valid, international and organisational setting.

Additionally, most studies so far have investigated links between resilience/well-being, coaching and performance by utilising cross-sectional designs and subjective self-report measures that may obscure the actual benefits of coaching (Dahling et al., 2016; De Haan, 2021). Furthermore, randomised controlled trials (RCTs), despite being the 'gold standard in clinical research', cannot be blinded to either coach or coachee. Nevertheless, the 40 or so randomised controlled studies we have now, even if mostly quite small scale and in academic as opposed to realistic business contexts, are the best evidence for the general outcome of workplace and executive coaching that we have (De Haan & Nilsson, 2023). They have been shown to demonstrate consistent effect sizes different from zero, and regular variation that appears to conform with a 'leverage' or 'co-regulation' model of coaching that suggests that the successful outcomes of coaching are associated with the greatest mutual influence between coach and coachee.

Given the limited number of rigorous studies to date, expanding the database of randomised controlled trials in executive and workplace coaching is essential, not only to enhance the realism of such research but also to enable direct comparisons between different coaching modalities. RCTs are the best means to collect evidence of an intervention which can be attributed to that intervention itself, and therefore the best way to check if the increasing corporate expenditure on coaching is warranted. They can also

provide clues as to the active ingredients (the independent variables that significantly interact with coaching outcomes) in coaching. RCTs can show how such ingredients conspire to moderate or mediate the effects of coaching, helping to maximise investment yields in coaching. In general, RCT designs can help to find out which individual coachee might benefit most from the intervention, or which preparation or coaching approach, will be optimal. Currently, there is just one RCT involving AI coaching (Terblanche et al., 2022), but unfortunately, it does not make a direct comparison with human coaching. The current study is specifically designed to expand on current research by making more direct comparisons between human and AI coaches and including multiple new outcome measures.

Evidence from RCTs can help to improve the coaching interventions and clarify which techniques or types of contracts give the highest return on investment. Many organisations are forced to be selective in terms of both the executives and managers that may apply for workplace coaching and the (internal or external) coaches that can provide this service. Costs are the main driver behind the recent growth of AI chatbots (Terblanche, 2020). Thus, the results of RCTs can play an important role in making delicate choices about investment in coaching, which type of coaching and the matching of coachees and coaches. For example, Page and De Haan (2014) suggest, on the grounds of outcome research to date, that it is better not to undertake structured and systematic matching based on demographics or personality fit. Contrary to popular recommendations, current findings suggest that random matching and self-selection based on initial 'alignment' or 'chemistry' sessions determined by the prospective coachees themselves, yield the best coaching outcomes (Page & De Haan, 2014).

In this non-preregistered study, we report on a coaching RCT study in an industry-based setting, within a mid-sized global corporation in education, comparing and building on initial results from AI coaching and human coaching directly (Terblanche et al., 2022). The study contributes to the existing coaching outcome literature in four important ways:

(1) It partially replicates and builds on earlier AI digital coaching findings (Terblanche et al., 2022);
(2) It studies the outcome of executive coaching in a naturalistic corporate setting with coachees who are senior leaders and coaches who are experienced and accredited executive coaches, whilst still studying coaching uncontaminated by other leadership-development activities;
(3) This study demonstrates the impact of coaching on realistic outcomes, such as success on coachee-stated goals and validated wellbeing indicators, which are both psychometric and physiological;
(4) By being able to follow up with all coachees in the study due to having no attrition in the response to questionnaires, we can disambiguate the effects of poor adherence and study attrition.

This design is provided for tracking the uptake of coaching during the entire study which lasted approximately 3 months (Figure 1): time point T1 is days before coach and coachee are matched, T2 is after the experimental groups have received human and AI coaching, respectively and T3 is at the end of the experiment, when the waiting-list control group

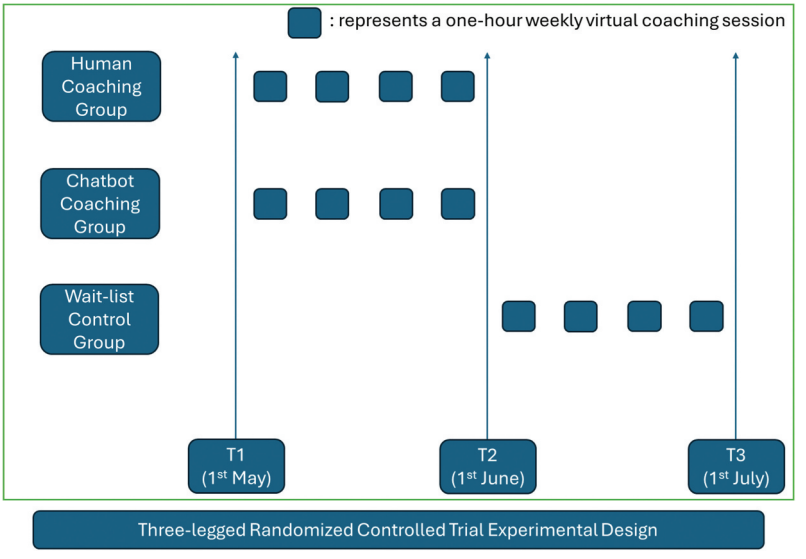

Figure 1. Experimental design of the three-legged randomised controlled trial. Please note the human coaching and waitlist control groups had agreed sessions of 1 hour weekly, whilst the chatbot coaching group was requested to consult the coaching chatbot once a week for 1 hour.

has also received human coaching. The present study is an expansion on previous research in the following ways: 1. The comparison between human and AI coaching is direct, using the same control group; 2. the chatbot is enhanced with generative AI (see Methodology section); 3. the subjects are senior managers as opposed to students in the first experiment; and 4. the frequency of sessions is higher but duration the same (1 h per session).

Theory development: co-regulation in coaching

The coaching model used for this study is based on the *beneficence* of HRD interventions, where we assume that coaching is an opportunity and a luxury, and if motivation is high and the company pays, it is an attractive engagement costing only the time and energy spent in conversation. We assume, therefore, that the so-called 'negative therapeutic reaction' (Horney, 1936), which we know is rare even in psychotherapy, and mostly occurs in intractable cases and long collaborations of over 20 sessions, does not play a role in outcome (i.e. coaching outcomes will be mostly desirable and positive for the coachee). Assuming the coaching offer is in principle high-status, desirable and voluntary (e.g. in our experiment there was no cost to the organisation, no cost to the coachee and coachees have stepped forward of their own free will and initiative), it is expected that there will not be a negative reaction. Any disappointment could simply be resolved by walking away or changing the coach.[1]

Our experimental design predicts that coachees can make good use of their initial motivation, well-being or resilience and that attrition is not a sign of a negative impact of coaching ('dropout') but rather the recognition that the coachee is better off regarding her/his goals by not continuing in the coaching. Goal-setting theory demonstrates that

setting specific, challenging but achievable goals enhances resilience and self-efficacy. Moreover, a specific goal setting predicts resource development (learning), performance and satisfaction, with these effects moderated by participation, commitment and feedback (Locke & Latham, 2002). These benefits can be further amplified through clear communication, regular feedback and by documenting goals, as was done in this study.

If coach and coachee are motivated to achieve coaching goals, they will both work actively towards achieving those results in the time allocated. This means they will collaborate and *co-regulate* (Tschacher et al., 2014) to achieve the greatest results, and their coaching outcome will be determined by their (mutual and individual) *leverage* or influence on obtaining results (i.e. if they can take different decisions, implement changes, or achieve other objectives achieved within sessions, they will freely undertake to do so). This leads to our assumption that 'common factors,' including coach credibility and motivation, coachee self-efficacy and motivation, coachee seniority in the organisation and trust and bond within the coaching alliance, will matter most to the outcome because they increase leverage. There is already considerable evidence to support each of these individual factors contributing to the overall coaching effectiveness (De Haan & Nilsson, 2023).

Similarly, one would expect the number of sessions to be uncorrelated to the outcome over a broad range, as coach and coachee will 'fit in' the work to the sessions allocated. This has already been found in some studies (Anthony et al., 2013; Goff et al., 2014) as well as several meta-analysis studies (De Haan & Nilsson, 2023; Sonesh et al., 2015; Theeboom et al., 2014).

Dropout during the coaching, which leads to attrition, can also be understood as another freedom the coachee has to optimise the coaching, namely the freedom to exit at the optimum moment. As a result, our model would predict that dropouts are equally transformed on average, as do those who stay in the coaching.

A lot less is known about what makes AI coaching successful, but the current thinking is that *the leverage* that coachees have on the chatbot is paramount, including factors such as the ease of use, openness to a variety of topics and goals, perceived relevance and meaningfulness of chatbot responses. One RCT study on chatbot adoption (Terblanche & Kidd, 2022) investigated the technology adoption factors that influence coachees' decision to use an AI-chatbot coach. They found that performance expectancy (does the chatbot do what it claims?'), social influence (credibility of the chatbot, or 'what others think about me using it?') and facilitating conditions (being able to use the necessary equipment) played a significant role in coachees' decision to use the chatbot.

Age and gender also played a role (e.g. older people were more concerned with the effort involved). Other studies have looked at the role of modalities such as text versus voice (Terblanche et al., 2023), adding images to chatbot text prompts (Terblanche & Prywes, 2025) and clicking versus typing when interacting with the chatbot (Mai et al., 2022). The conclusion at this stage of coaching research is that it is not yet clear exactly what type of AI coaching chatbots appeal to which coachees but that the leverage that coachees feel they have is important, like human coaching.

Study hypotheses

Based on the effect sizes found in previous studies, we expect human coaching to be demonstrably effective with a somewhat smaller effect size than in psychotherapy, where

effect size *d* currently stands slightly above 0.7 (Cuijpers et al., 2020). Based on earlier randomised controlled trials in coaching (Barger, 2025; Terblanche et al., 2022) we would predict that AI coaching will likely have a significant impact on outcome scores as well, but to a lesser extent on individual factors, such as wellbeing or self-efficacy.

H1: Coaching outcome scores (a) with human coaches are significantly different from the control group, with standard effect sizes between 0.5 and 0.7, on all outcome measurements; (b) however, with AI-chatbot coaches, the outcomes are significantly different between coachee and control groups only when it comes to goal success.

Evidence for a number of different 'readiness' factors predicting overall coaching effectiveness and success outcomes have been found in previous studies, including psychological capital, openness as one of the Big 5 personality factors and coach-related skills. Psychological capital (e.g. Dyrbye et al., 2019; Fontes & Dello Russo, 2021; McGonagle et al., 2014, 2020) is a combined measure of hope, optimism, self-efficacy and resilience that moderately correlate with one another, suggesting potential confounds in the conceptualisation and measurement of these constructs or pointing to perhaps a single underlying factor. We define this intercorrelated set of dimensions, for convenience, as 'readiness' coaching factors. The Working Alliance Inventory can also be considered a 'readiness' factor, as it is a well-established measure of the strength of the coach–coachee relationship (Horvath & Greenberg, 1986) and has been shown to predict coaching effectiveness in a broad, general sense (e.g. Graßmann et al., 2019), rather than as a relational variable that determines outcomes from session to session (De Haan et al., 2020). We expect predictive power to be shared by all 'readiness' factors and none of these factors to mediate any of the others.

H2a: The coaching outcomes in terms of goal success scores of all cohorts are predicted by 'readiness' factors, such as General Self-Efficacy, Hope, Resilience/Hardiness, Wellbeing, Perceived Social Support, (inversely by) Perceived Stress and by the perceived ability to build relationships which can be expressed by Working Alliance. H2b: All these 'readiness' scores should also predict coaching effectiveness as scored by coaches, goal success as scored by line managers, and momentary measurements of their cardiac rhythms.

Previous RCT studies have lost significant participation of coachees through *attrition* related to both the coaching engagement and the questionnaire completion. Attrition has been found to be relatively high in coaching, with attrition in the 39 RCT studies of De Haan and Nilsson (2023), on average 14%, with a maximum of 55% in one study. Additionally, among the 39 there are 12 coaching RCT studies with an attrition of 0%. However, this is achieved through replacement of a pipeline of candidates or a high control of the subjects through other dependent relationships (students, registrars, employees, paid subjects), which obscures the real effect of attrition on results. In these studies, dropouts have become absent and hidden even if their disappearance may well be due to coaching-related adverse effects. It is rare to find adequate reporting of dropout and attrition in coaching reported in RCTs. Nevertheless, at least two of the published experiments in coaching failed in their randomisation due to massive attrition (Egan & Song, 2005; Ungerer et al., 2019).

A major critique that can be levelled at earlier RCT studies is therefore that, by allowing attrition, they were only studying those coachees that *did* benefit from coaching (e.g. De Haan, 2021). Arguably, those that *did not* benefit from coaching, and those that were affected adversely by coaching (i.e. experiencing negative outcomes), will have left the experiment before the final measurement, and are therefore not taken into the statistics. Moreover, if the attrition rate is as high as it seems to be in coaching, there is a second concern, namely that perhaps coaching cannot reach the full spectrum of participants (executives) that it purports to help. Ward et al. (2017) suggest that attrition (and, similarly, careless responses) are major sources of bias in online surveys, which can be related to the subject's personality profile.

In this study design, we effectively retrieved all dropouts: everyone who started with the first measurement will complete the final questionnaire, simply because we remind all of them (by email, text, app and then even phone) until and unless they eventually complete their next questionnaire. This was done in preservation of informed consent, due to the fact that all leaders were generally motivated to help and were remote colleagues of the first author, so they respected his personal pleas.

A unique and important novel feature in this study was therefore that we were able to monitor engagement with coaching independently of the completion of questionnaires. If there was non-engagement with human coaching, the coaches would tell us (Note: this happened in just three cases and in one case there was a language issue, so we asked a Spanish-speaking coach to take over from the originally assigned English-speaking coach). Likewise, in the AI group, the chatbot software would signal that the coaching had not been completed. One coachee told us that they were simply 'too busy' to engage with any of the coaching but 100% of the questionnaires were still collected from them.

A total of 114 coachees were followed throughout the study, but only 99 coachees completed the experiment in the intended way. This means our attrition amounted to 13%, which is within the 10–15% range normally found in RCTs in the helping professions. When attrition is studied in detail in medicine, it generally does not seem to alter the main outcomes of the experiment and was demonstrated to act like a random process (Crutzen et al., 2013; Hewitt et al., 2010). However, attrition through dropout in psychotherapy can be much larger (more than twice the level found here and in medicine generally) and to a large degree related to negative outcomes (e.g. in-person therapy: Dixon & Linardon, 2020; smartphone app-based therapy: Linardon & Fuller-Tyszkiewicz, 2020) and an overestimation bias. As such, attrition can be neutral to the experiment, but it can also be a serious bias, particularly in the helping professions and beyond (Millán & Macours, 2017).

In this study, all participants were drawn in by the promise of coaching and sought no other rewards. Four of those who dropped out from the AI coaching requested human coaching instead and could be moved to a waiting-list control group, which is a strong sign of them wanting to be coached (we still count these as evidence of attrition, because they did not finish the experiment in the intended way). We expect attrition to not fundamentally change the results on effectiveness in this study, which was the first RCT in coaching where this could be tested independently.

H3: The randomised controlled trial's main results are not affected by attrition: even though those subjects who have not taken much from the coaching are predicted to be more likely to attrit, leaving them out of this experiment will not affect the results measurably.

Methodology

The current study involved 114 leaders in a private multinational business in education, of various levels of seniority. All business leaders in the organisation were offered free coaching by experienced and qualified coaches in the month of May 2024, consisting of four weekly one-hour sessions with an accredited coach. After the application, in late April (T1; see Figure 1), a baseline questionnaire was sent out and the coachees were randomly allocated to three equal-sized groups: traditional human coaching, AI-chatbot coaching and a waiting-list control group which was offered the same 4-h human-coaching experience in the month of June.

By asking coachees, coaches and line managers to rate outcomes on overlapping scales at T2 (end of May) and T3 (early July), we have three independent measures of coaching outcomes. We had 74 different executive coaches working with 78 coachees, so we avoided a multilevel data structure, preventing the statistical complications of nested data. We checked for the influence of gender, age and ethnicity (the coachees originated and were based in nearly 30 different countries, adding to the internationalisation and generalisability of findings from this study) and rank in the organisation.

Research organization, approval, confidentiality and process

The organisation in which the study was situated is a global corporate in the educational sector, comprising business and language schools, with approximately 50,000 employees based in over 110 countries. Recruitment was also open to line managers at one of the authors' other universities, again through senior management and HR, which yielded 6 more university leaders. Everyone was offered four sessions of 1-hour, Zoom (face-to-face human) or WhatsApp (text-based AI) coaching for the first four Fridays of the month, either with an AI chatbot or with an experienced and accredited executive coach during either May 2024 or June 2024. The study was granted research board review approval at the first author's first institution (Reference #ID-RP-326). Adhering to the ethical guidelines meant obtaining permission from all coaches and coachees to join the study via informed consent and ensuring full confidentiality of participants by storing only numerical identifiers.

Based on an a *priori* power analysis starting from the recent RCT-studies' meta-analysis (De Haan & Nilsson, 2023), we expected to measure an approximate effect size of d = 0.70 for goal attainment, which falls exactly in the middle of the g = 0.59 found for all 40 earlier RCT's and the g'= 0.82 found more specifically for the goal attainment variable in 10 out of those 40 studies. Assuming such a medium effect size (Cohen's d = 0.70), 0.80 statistical power and a 0.05 significance level (two-tailed), 34 participants per group were needed per experimental group (i.e. $3 \times 34 = 102$ in total).

Table 1. Assessments by experimental groups and time. Note: manager data were only collected if the coachee had given permission.

Study variables	Coachee T1	Coachee T2 + T3	Coach T2 or T3	Line manager T1 + T2 + T3
Demographics: Gender, Age, Nationality, Country of residence, Ethnicity and Rank/Experience level	*			
General Self-Efficacy scale (total)	*	*		
Hope scale (total)	*			
Brief cognitive hardiness scale (total)	*	*		
Perceived social support scale (total)	*			
Perceived stress scale (total)	*	*		
Psychological wellbeing scale (total)	*			
Goal 1: Success	*	*		*
Goal 1: Difficulty	*	*		
Goal 2: Success	*	*		*
Goal 2: Difficulty	*	*		
Outcome Expectations		*	*	
Coaching Effectiveness (CE)		*	*	
Working Alliance		*		

Before being told which group they belonged to, all coachees completed the first of three research questionnaires (the first one was sent on 19 April 2024 or T1), and after completion they were allocated a coach if in the first two experimental groups. The second questionnaire was sent out after the RCT experiment, on 31 May 2024 (T2), after which subjects in the waiting-list control group were allocated to their coach. The third follow-up questionnaire was sent out after the control group had received coaching, on 5 July 2024 (T3). Coaches received a questionnaire at the end of their engagement, so on 31 May and on 5 July, for the two cohorts of coaches. Line managers received the three requests to rate their reports on their coaching goals after completion of the coachee's questionnaire and only when the coachee had given permission for the line manager to rate them. Figure 1 summarises the study design and the content of each of the three questionnaires per cohort is provided in Table 1.

In contrast to most other randomised trials in this field, we decided to keep those subjects who did not take up the coaching inside the sample and to address attrition by insisting that everyone would complete each questionnaire. We sent as many emails and text messages as were needed to get all the questionnaires completed. By means of this careful follow-up, we were able to control our attrition levels. All coaches, coachees and managers responded to the questionnaires at T1 and T2. Only at the final hurdle, at T3, two line managers dropped out because they left the company. We believe that this is an important design element, as non-uptake or only sporadic uptake of coaching (i.e. non-commitment) weighs on the general effectiveness of coaching and is one of the reasons why effectiveness may be lower than reported. By having all questionnaires completed, we can analyse both with and without completion of the coaching – whilst other studies (cf., those in De Haan & Nilsson, 2023) have generally confounded attrition of survey completion and non-uptake of coaching.

Subjects

Coachee information

Top managers and HR leaders from the two organisations posted emails to all line managers locally, followed by one or two reminders, starting 26 March, and until late

April 2024. A total of 120 middle and senior managers were recruited but there were six subjects who declined to participate in the study after randomisation and before the coach was allocated, as they felt they would not have enough time for it. For this reason, our study's design included 38 participants in the human-coaching group, 37 in the AI-coaching group and 42 in the waiting-list control group. Those in the control group received the exact same human-coaching offer as the first group, exactly 1 month later (June instead of May). Average age of the subjects was 39.1 years, 67% female, 32% male, 1% trans, with 47% at 'director' level, 39% 'manager' level and 13% 'project leaders. Subjects came from 34 countries all over the world (13 ethnicities; 69% white) so the sample was diverse and representative in nature.

Coach information

The externally recruited coaches were experienced (the average experience level was 14.5 years of practice) and all shared the level of coaching competence that was the minimum level for being on the coach list of a major business school. This included a minimum of 5 years' experience in executive coaching, evidence for regular supervision and continuous professional development and a minimum level of accreditation (ICF PCC/MCC or Ashridge accreditation). The coaches were allocated randomly to their coachees and offered 4 hours of their time for the four weekly coaching sessions. At the conclusion of the study, the coaches were rewarded with a preview of the experiment's results (after all data collection) and a webinar on the methodology and uniqueness of this trial. In total, 74 coaches participated in the study, with an average age of 56.1 years; 53% were women, representing 28 countries and a wide range of ethnic backgrounds, of which 79% were white. All coaches were randomly assigned to their respective coachees. Only four coaches took on a second coachee in the experiment to help with the control group.

AI coach and development of the platform

The AI coach used in this study was a custom-developed, text-based chatbot accessed via WhatsApp, employing a hybrid scripted and generative AI design (Terblanche & Tau, 2024). In general, two types of AI coaches are used in coaching practice (scripted and generative). In scripted chatbots, the conversation flow is predetermined using a finite decision tree that determines the direction of the conversation. Chatbots using generative AI, on the other hand, rely on large language models (LLMs) like ChatGPT to produce conversation output based on prompts (Terblanche, 2020). The unique hybrid AI chatbot used in the present study used a scripted framework to ensure adherence to goal theory and reflective practice, but infuses the scripted conversation with generative AI to provide context-aware questions. There is, therefore, not a single prompt that captures the entire coaching chatbot design, but rather a series of short prompts to construct a question or summary based on the conversation history.

Overall, the design was based on the Designing AI Coaches (DAIC) framework (Terblanche, 2020). The DAIC framework merges human coaching efficacy aspects of a strong coach–coachee relationship with chatbot design best practices. It recommends that, due to the existing limitation of ANI, the AI coach should be designed to fulfil one specific coaching task only, in line with expert system design principles (Terblanche, 2020). This recommendation was used to inform the design of the chatbot to have the sole purpose of being a goal-attainment chatbot coach. The choice

of goal attainment as the focus for the chatbot was based on the fact that while coaching has numerous positive outcomes, goal attainment is arguably what distinguishes coaching from other similar helping professions (Grant, 2008, 2014; Locke & Latham, 2002).

Line manager information

The line managers participating in this study directly supervised the 81 coachees (71% of the total sample) who had given their supervisor permission to know about their goals and rate them before, during and after the experiment. Those subjects gave us the line managers' email address (or in one case two email addresses when a different line manager was responsible for each of the two coaching goals). This email address was used to send an email with a statement of both goals of the coachee each time after the coachee had completed a research questionnaire, i.e. three times in total (T1: early May; T2: late May; T3: early July)–and, as with the coachees, enough follow-up emails were sent to achieve a 100% response rate for all line managers who stayed with the company (79 out of 81 stayed during the entire 10-week period; i.e. some attrition for managers, but none for coachees).

Study measures

Given the brief 3-month design of this comparative coaching study, no meaningful work performance measures could be collected. Consistent with many previous executive-coaching studies (De Haan, 2021), several established and validated instruments were included to measure effectiveness, work goal success, resilience factors and wellbeing outcomes. All these measures have established psychometrics and have been used in a wide range of outcome-based coaching studies.

Coaching effectiveness (CE)

CE was assessed using four items on a 7-point response scale (ranging 1 = Strongly Disagree, Disagree, Slightly Disagree, Neither Agree nor Disagree, Slightly Agree, Agree, to 7 = Strongly Agree) and asking for the following: 'The outcome of my coaching objectives so far: (1) I have been successful in creating reflective space for me; (2) I have been successful in creating new insight for me; (3) Through (preparation for) coaching I have successfully engaged in new action or behaviour; (4) I would consider this coaching journey successful'. We chose this scale because it has been used in prior coaching publications and was shown to have good reliability (De Haan et al., 2013, 2016, 2019, 2020; Kuntz & Livingston, 2020) and because it seems to map a wide range of areas of possible effectiveness of coaching. Responses were calculated as the average score across these four items and Cronbach's Alphas are found in Table 2.

Coach-coachee relationship

The short-form Working Alliance Inventory (WAI) in short form by Tracey and Kokotovic (1989) was used as a measure of the strength of the coach–coachee relationship (Horvath & Greenberg, 1986). The WAI measures three specific dimensions, scored on a scale of 1–7, with 1 being 'never' and 7 being 'always' with higher overall scores indicating stronger alliance:

Table 2. Pearson correlation matrix and Cronbach's alphas of study variables.

Variables	1	2	3	4	5	6	7	8	9	10
(1) General Self-Efficacy scale (total)	0.75									
(2) Adult Hope scale (total)[¥]	0.42***	0.83								
(3) Brief cognitive hardiness scale (total)	0.41***	0.55***	0.79							
(4) Perceived social support[¥]	0.15	0.18	0.23*	0.93						
(5) Perceived stress scale (total)	−0.32***	−0.32***	−0.58***	−0.12	0.85					
(6) Mental Wellbeing scale (total)[¥]	0.35***	0.46***	0.62***	0.42***	−0.69***	0.91				
(7) Goal 1: Success: Self-rating	0.29***		0.35***		−0.37***		1			
(8) Goal 1: Difficulty: Self-rating	−0.03		−0.07		0.18**		−0.28***	1		
(9) Goal 2: Success: Self-rating	0.29***		0.35***		−0.28***		0.65***	−0.05	1	
(10) Goal 2: Difficulty: Self-rating	−0.11*		−0.09		0.1		−0.06	0.29***	−0.21***	1
N	333									

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, [¥]Only measurements for T1 ($N = 114$), Cronbach's are in the diagonal.

- *Task* refers to what coach and coachee agree needs to be done for the coachee to reach his/her goals for coaching. A sample item is 'I am clear as to what my coach wants me to do in these sessions'.
- *Goal* refers to the outcomes that the coach/coachee hopes to gain from coaching. A sample item is 'the goals of these sessions are important to me'.
- *Bond* refers to which extent the coach/coachee trusts, respects and feels confidence in the other person. A sample item is 'I believe my coach is genuinely concerned for my welfare'.

General self-efficacy

The General Self-Efficacy scale was used to assess the coach/coachee self-efficacy (GSE; Schwarzer & Jerusalem, 1995). This scale consists of 10 items on a 4-point Likert scale. Sample items include: 'I can always manage to solve difficult problems if I try hard enough', 'If someone opposes me, I can find the means and ways to get what I want' and 'It is easy for me to stick to my aims and accomplish my goals'.

Hope

Hope was measured for all coachees using the Adult Hope Scale (Snyder et al., 1991). This consisted of 12 items (sample item: 'There are lots of ways around any problem'), which separate into two 4-item subscales measuring the underlying dimensions of Agency (i.e. goal-directed energy) and Pathways (i.e. planning to accomplish goals) and four further 'filler' items, which we included in the survey but discarded for measurement purposes. The response coding for each item ranged from 1 (definitely false) to 8 (definitely true).

Cognitive hardiness

Cognitive hardiness was measured using the brief 15-item (α = 0.79) Cognitive Hardiness scale (Nowack, 1990). Example items are: 'When all else is bleak, I can always turn to my family and friends for help and support' and 'I expect some things to go wrong now and then, but there is little doubt in my mind that I can effectively cope with just about anything that comes my way' rated on a 5-point agreement scale from 1 (strongly agree) to 5 (strongly disagree). This measure was selected because it has been utilised in several previous randomised (waiting-list) controlled coaching studies with significant changes from pre and post measures (Grant, 2008, 2014).

Perceived social support

Social support was measured with Zimet's Multidimensional Scale of Perceived Social Support (MSPSS; Zimet et al., 1988). The scale consists of 12 items on a 7-point Likert scale, mapping three subscales related to support from Family, Friends and Significant Others, with sample items 'my family really tries to help me', 'my friends really try to help me' and 'there is a special person who is around when I am in need'.

Perceived stress

Perceived stress was measured using the 4-item short-form of the Perceived Stress Scale of Cohen et al. (1983), which measures the degree to which a person evaluates situations in his or her life as stressful. A sample item is, 'In the last month, how often have you felt that you were unable to control the important things in your life?' and the response coding for all items ranged from 1 (*never*) to 5 (*very often*).

Mental wellbeing

Mental Wellbeing was assessed using the 14-item Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS; Tennant et al., 2007), which covers both psychological and subjective aspects of satisfaction and happiness, or technically 'eudemonic' and 'hedonistic' criteria of Mental Wellbeing.

Neurophysiological

Momentary cardiac activity was measured using a commercial platform (Immersion Neuroscience, Henderson, NV; www.getimmersion.com). Neurologic Immersion measures the value the brain obtains from social-emotional experiences, such as coaching by applying algorithms to heart-rate data (Merritt & Zak, 2024). Immersion has two main components: attention to the experience one is having, associated with dopamine binding to the prefrontal cortex and the emotional resonance of the experience associated with oxytocin release from the brainstem (Immersion and Safety).

The Immersion Neuroscience platform specifically captures neuroelectric activity associated with dopamine and oxytocin in the cranial nerves using a photoplethysmogram (PPG) sensor and algorithms applied to cardiac data. These signals are convolved into a single measure called Immersion that was designed to accurately and consistently predict behaviours (Lin et al., 2022). The data were sent to the cloud continuously via subjects' mobile phones and were also viewable in real-time. The Immersion Neuroscience platform provided an output file used in the analysis and no native data from subjects' wearables

were used to build parsimonious models. We chose to measure neurologic outcomes for this coaching study because of the well-established relationship between social interactions and positive mood states (Merritt & Zak, 2024; Merritt et al., 2022). Moreover, by accessing neurophysiologic signals from smartwatches, continuous nonintrusive data was obtained from subjects without affecting their daily activities.

Data pre-analysis checks

We tested the differences between the three random groupings on 16 different dimensions measured at T1 (including managers' feedback on goals), by means of t-tests and we found none. Descriptive statistics, reliability estimates and inter-correlations for all time points are reported in Table 2. As in other studies employing the same variables, all variables show a good internal consistency ($\alpha > .78$ throughout; see Table 2). We also checked for any differences between male and female coachees as well as for age, ethnicity and seniority in the organisation and found no significant results indicating there were no biases in our sample. We found some convergence in outcome variables measured by different individuals, namely Coach Effectiveness: between coaches' and coachees' scores (correlation 0.49***) and the same between coachee and line-manager goal success: correlations were 0.18 for Goal 1 success and 0.29* for Goal 2 success. The chatbot use was an average of 4.79 sessions with a mean duration of 17 min 55 s over all sessions.

Results

Hypothesis 1: effectiveness of coaching

To ascertain the main result of the randomised controlled trial we undertook a two-sided t-test for all key outcome variables between the three experimental groups (see Table 3). We also conducted linear mixed-model regressions that confirmed all results in Table 3 (materials can be requested from the first author).

Across most outcome measures, the experimental group that has worked with human coaches scores significantly different from those in the control group, supporting Hypothesis 1a. On the other hand, those that worked with AI-chatbot coaches did not score differently from those in the control group on most variables (see Table 3).

Thus, we do not find support for Hypothesis 1b. Even if we regress the AI results on the time spent with the chatbot, we find that results only improve slightly but that differences between Control group and AI group stay below the .05 significance level (including the result for General Self-Efficacy in Table 3 which does not become more significant). Time spent on the chatbot does not correlate with any of the scores at time T2, except for perceived stress where it was 0.40*, indicating that possibly less stressed individuals spent more time on the chatbot.

Standard effect sizes on realistic variables are at least 0.5, representing a medium effect size for self-efficacy and goal success self-scores. Coaching effectiveness and WAI are unsuitable comparison variables because the items refer to coaching and the coaching relationship which the control group has not undertaken at T2. The 'post' measurement at T3, after the waiting-list control group was coached, shows the same picture. Improvements for the human-coaching group and the (by T3 also human-coached)

Table 3. Two-sided t-tests of key variables between the experimental coaching group and the control group at time T2.

Variable	Control Group $n = 39$	Human Coaching group $n = 38$	AI Coaching group $n = 37$	Difference Control & Human	p-value	Cohen's d	Difference Control & AI	p-value	Cohen's d
General Self-Efficacy scale (total)	18.795	20.395	20.378	1.600*	0.012	0.584	1.584**	0.005	0.662
Brief cognitive hardiness scale (total)	56.846	58.526	57.351	1.680	0.251	0.263	0.505	0.734	0.078
Perceived stress scale (total)	25.077	22.421	25.135	−2.656*	0.025	−0.523	0.058	0.962	0.011
Goal 1: Success: Self-rating	0.497	0.653	0.551	−0.155**	0.007	0.633	0.054	0.388	0.200
Goal 1: Difficulty: Self-rating	3.179	3.184	3.081	0.005	0.968	0.009	−0.098	0.464	−0.171
Goal 2: Success: Self-rating	0.459	0.647	0.516	0.188***	0.001	0.753	0.057	0.348	0.217
Goal 2: Difficulty: Self-rating	3.308	3.026	3.243	−0.281	0.061	−0.436	−0.064	0.617	−0.115
WAI (total)	34.897	66.895	40.054	31.997***	0.000	1.951	5.157	0.243	0.268
Coaching effectiveness scale (total)	18.897	31.105	19.730	12.208***	0.000	2.922	0.832	0.512	0.152

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

control group, but the AI group very much lagging, as it did not improve significantly on any of the outcome variables. See Figure 1 for an overview of results at all points on our primary outcome variable: improvement in self-stated goal success, and on the combined variable: goal attainment scores (the average of goal success on both goals multiplied by their difficulty) (De Haan et al., 2019, 2020; Grant, 2008), measured as a standardised effect size starting from the T1-measurements. Only the human-group in T2 and the control group in T3 show significant effect size jumps ($p < 0.001$ and $p < 0.05$, respectively, for Goal 1 success). Overall, these results support previous meta-analytic findings. Wang et al. (2021) found that psychologically informed coaching approaches facilitated effective outcomes also in goal attainment (g = 1.29) and self-efficacy (g = 0.59).

There are two exceptions which could be statistical fluctuations (see Table 3): cognitive hardiness does not significantly differ between the human-coaching and control groups at T2 and T3, and General Self-Efficacy does stand out positively for the AI-coaching group at both time points. However, when looked at analyses within the sample we do see cognitive hardiness significantly improving in the human-coaching group (d = 0.45; $p = 0.05$ between T1 and T2) whilst the small growth in General Self-Efficacy in the AI-coaching group becomes clearly insignificant (d = 0.05; $p = 0.82$ between T1 and T2). Similarly, upon regression with the control variables, the cognitive hardiness scale remains significantly different for the human-coaching group, while the effect of self-efficacy in the AI-group disappears.

Hypothesis 2: 'readiness' factors predict goal success and coaching effectiveness

We checked predictions of outcome scores from the T1 independent variables (self-efficacy, hope, cognitive hardiness, perceived social support and perceived stress) and we added to these the perceived working alliance scores (WAI) at T2, because they have earlier been found to predict outcomes like the 'psychological readiness' factors (De Haan et al., 2020). Table 4 shows the regression for T2 onto Goal 1 success, which shows that the independent variables together explain 31.8% of the variability in Goal 1 success. This measure changes to 27.5% for regression onto Goal 2 success at T2, still very substantial, supporting Hypothesis 2a.

The prediction of the outcomes at T3 is very similar: the independent variables at T1 and T2 (WAI) predict 32.8% of the variability in Goal 1 success at T3, and 18.2% of the variability in Goal 2 success at T3. The only variable that does not reach significant correlation with these outcome variables is Perceived Social Support. Adult Hope turns out to be the best predictor of goal success, with correlations 0.48*** and 0.38*** for both goals. Perceived Stress is a strong second with (as expected: negative) correlations 0.33** and 0.25** for both goals. Differences between the individual predictions are not significant and, as we can see in Table 2 all the 'readiness' scales intercorrelate strongly at every time measurement. Proposed active ingredients, such as coaching relationship and coachee 'readiness', are therefore confirmed.

Regarding hypothesis 2b, we have also checked for the other, more objective measurements of outcome (line-manager rated goal success; coach-rated coaching effectiveness; Immersion and Safety scores) and despite the lower statistical power of these variables (with N, respectively, 81, 78, 14), we found some significant correlations. In fact, the same strongest predictors of the self-scored outcomes reached significance again in some cases: Adult Hope 0.73** with Immersion, Perceived Stress with Immersion 0.57*, with coach-rated outcome −0.26* and with manager-rated Goal success −0.34*. However, Perceived

Table 4. Regression results for the whole sample predict the goal success at T2. Note: the dependent variable is Goal 1 success at T2.

DV: Goal 1 Success at T2	(1) coef/se	(2) coef/se	(3) coef/se	(4) coef/se	(5) coef/se	(6) coef/se	(7) coef/se
General Self-Efficacy scale (total)	0.023*						0.002
	(0.010)						(0.011)
Adult Hope scale (total)		0.015***					0.011*
		(0.004)					(0.005)
Brief Cognitive hardiness scale (total)			0.009**				−0.005
			(0.003)				(0.005)
Perceived Social Support scale (total)				0.009			−0.002
				(0.008)			(0.009)
Perceived stress scale (total)					−0.013**		−0.011
					(0.004)		(0.006)
Mental Wellbeing scale (total)						0.010**	0.001
						(0.003)	(0.006)
WAI (total) at T2							0.005***
							(0.001)
Constant	0.118	−0.245	0.044	0.418**	0.773***	0.061	0.115
	(0.205)	(0.202)	(0.192)	(0.136)	(0.067)	(0.170)	(0.365)
Number of observations	114	114	114	114	114	114	114
R^2	0.049	0.127	0.059	0.010	0.095	0.084	0.318
Adjusted R^2	0.041	0.119	0.051	0.001	0.087	0.075	0.273

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; robust standard errors in parentheses.

Social Support correlates negatively with the coach-rated outcome (−0.26*). These are reassuring findings for a relatively small RCT, e.g. one much larger-scale experiment found similarly significant coach-rated prediction strengths (De Haan et al., 2016). Hypothesis 2b is therefore somewhat supported as well.

Statistical analysis suggested that mediation was not indicated because readiness factors do not predict goal success strongly (Table 4). In other words, goal success is near-insignificant in all readiness variables (self-efficacy, hardiness, wellbeing and perceived stress).

Hypothesis 3: attrition does not impact on the RCT results

Fifteen subjects (an attrition of 13%) did not complete the RCT experiment as intended, although they did complete all their questionnaires. Attrition was clearly biased: one individual from the human-coaching group did not take up the coaching (as reported by the coach); nine individuals from the AI-coaching group told us that they had not taken it up (four of these requested a human coach for the next time period, i.e., T2 to T3, i.e. together with the waiting-list control group, which we granted and organised for them) and three further subjects showing zero time usage of the chatbot; finally, two people from the control group did not take up their coaching when offered at T2.

Table 5 summarises the results of the study, omitting this attrited group of subjects (i.e. those who made no or minimal use of coaching). A comparison with Table 3 shows that the conclusions from the experiment remain the same. As expected, there is a slight sharpening of the results with larger effect sizes for the 'cleaner' group containing only active and engaged subjects in the experiment, but this difference is insignificant and much smaller than the Δd of 0.18 reported in clinical applications in the meta-analysis by Linardon and Fuller-Tyszkiewicz (2020), so no evidence of an attrition bias.

Limitations in interpreting the neurophysiological results

The Immersion Neuroscience platform uses momentary variations in cardiac rhythms to infer neurophysiologic responses of the cranial nerves as a measure of the neural value of social experiences (Zak & Barraza, 2018). The Immersion Neuroscience platform infers neural states from the activity of the cranial nerves using the downstream effects of dopamine and oxytocin on cardiac rhythms (Immersion) and vagal tone associated with heart rate variability (Psychological Safety). These data were collected for individuals who had four or more well-defined significantly elevated Immersion experiences previously shown to be associated with overall mood and wellbeing (Merritt & Zak, 2024; Zak & Barraza, 2018).

Due to technical challenges, full datasets are available only for 14 subjects. Analyses (paired t-tests) were conducted during the coaching intervention, divided into two periods based on a midpoint date (2024–05–24), representing the first and second halves. of the study. Due to the small physiologic sample, we were unable to compare differences between the human- and AI-Chatbot groups, so the results reflect an overall measure across all coaching programmes. Results found that mean Immersion, but not Psychological Safety (PS) scores were significantly different between the beginning and towards the end of coaching ($t = -3.05$, $p = .013$ and $t = -.45$, $p = .67$, respectively).

Table 5. Results after attrition: two sided t-test of key variables between experimental coaching group and the control group at time T2, with all subjects that did not complete the coaching taken out.

Variable	Control Group ($N = 37$)	Human Coaching ($N = 37$)	AI Coaching ($N = 25$)	Difference Control & Human	*p*-value	Cohen's d	Difference Control & AI	*p*-value	Cohen's d
General Self-Efficacy scale (total)	18.757	20.472	20.520	1.649*	0.013	0.592	1.763**	0.005	0.703
Brief Cognitive hardiness scale (total)	56.865	58.583	58.000	1.730	0.255	0.267	1.135	0.523	0.167
Perceived stress scale (total)	24.892	22.194	24.560	−2.622*	0.030	−0.515	−0.332	0.811	−0.063
Goal 1: Success: Self-rating	0.492	0.667	0.604	0.170**	0.004	0.699	0.112	0.111	0.424
Goal 1: Difficulty: Self-rating	3.189	3.167	2.960	0.000	1.000	0.000	−0.229	0.147	−0.411
Goal 2: Success: Self-rating	0.446	0.653	0.512	0.208***	0.001	0.834	0.066	0.354	0.247
Goal 2: Difficulty: Self-rating	3.324	3.028	3.120	−0.270	0.076	−0.419	−0.204	0.156	−0.366
WAI (total)	34.216	67.889	43.120	33.514***	0.000	2.086	8.904	0.054	0.465
Coaching effectiveness scale (total)	18.757	31.417	20.000	12.595***	0.000	3.078	1.243	0.389	0.229

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Overall, even with the very small sample size and limited power to detect meaningful differences, these results add evidence of significant changes in some neurological states during coaching that should be cautiously interpreted.

It is typical in coaching studies to end up with relatively small sample sizes when measuring physiological data (e.g. 10–50 participants) due to numerous methodological and design challenges posed by fMRI, EEG and physiologic wearables (e.g. Jack et al., 2017). Rather than omitting our results based on inadequate and underpowered sample size, we hope our piloting of the collection of such neurophysiologic measures will encourage coaching and management researchers to include and replicate our limited findings in the future.

Discussion

This study used a three-arm RCT with an AI chatbot coach, a human coach and non-treatment through a waiting list, inside an international educational company. The rigorous coaching programme spanning approximately 1 month in duration fits into today's fast pace and the competitive company landscape that leaders find themselves in. The inclusion of multiple validated coachee, coach and line-manager measures provides concurrent and convergent validation of several factors that impact coaching effectiveness. Several important results of this human versus AI coaching comparison study are briefly highlighted below.

Confirmation that executive coaching by humans can be effective, while questions remain over the ability of AI to replace human coaching

Our results confirm other published meta-analytic RCT findings in support of human coaching but challenge the overall effectiveness of an AI-coaching chatbot with this leadership sample. Our findings do not support an earlier published study (Terblanche et al., 2022) that did find that AI helps significantly but with goal achievement only, using an identical measure of goal success (Table 3 and Figure 2). This time we only find a significant difference for the AI group on General Self-Efficacy, which is precisely what was not found in Terblanche et al. (2022). However, the self-efficacy difference disappears in regression.

One possible explanation could be that for goal attainment to show effect with AI coaching, more engagement and longer timelines were needed, as was the case in Terblanche et al. (2022). Another explanation could be a nocebo or outcome-expectation effect: the participants in the AI-coaching group knew others were seeing a human coach and might have felt they missed out. Barger (2025) conceived of a balanced three-legged, coach-blinded RCT experiment where 81 students were randomly allocated to a (simulated) AI coach, a human coach, or a control group. The 'AI coach' was in fact provided by human coaching but interfaced by an avatar and voice distortion. The only difference between the two experimental conditions was therefore the belief that they were coached by AI versus by a human coach. From this unique experiment, we can take that some differences in experience and feelings of missing out on human coaching made no difference in terms of coaching effectiveness. Barger's experiment would therefore indicate that the differences found were due to actual coaching, although it needs to be stressed that in our case the visual/auditory differences between human and AI coaching were vastly increased compared to Barger (2025). Indeed, it is obvious to users that chatbot conversation is very different from human conversation: even the newest chatbots are far too superficial and forgetful, failing to focus on what is really relevant in the relationship and failing to properly emulate real human meetings (Smith et al., 2025).

It is possible that busy executives and leaders are less inclined to use AI support anyway, which would account for the partial success with student samples in earlier research. Replication of our findings with other client groups might help qualify for whom AI coaching technology might be most beneficial to enhance performance, goal achievement and mental health.

Active ingredients such as coaching relationship and coachee 'readiness' are confirmed

We have found a significant prediction of outcomes by coachee-related factors, such as cognitive hardiness, self-efficacy, perceived social support and wellbeing, all factors indicating a good 'readiness' for the impact of coaching (extending earlier results by De Haan et al., 2013, 2016, 2019, 2020; Stewart et al., 2008; Zimmermann & Antoni, 2020).

The coaching relationship, as expressed by the perceived working alliance, also seems to have a predictive power confirming Graßmann et al. (2019). As these factors also

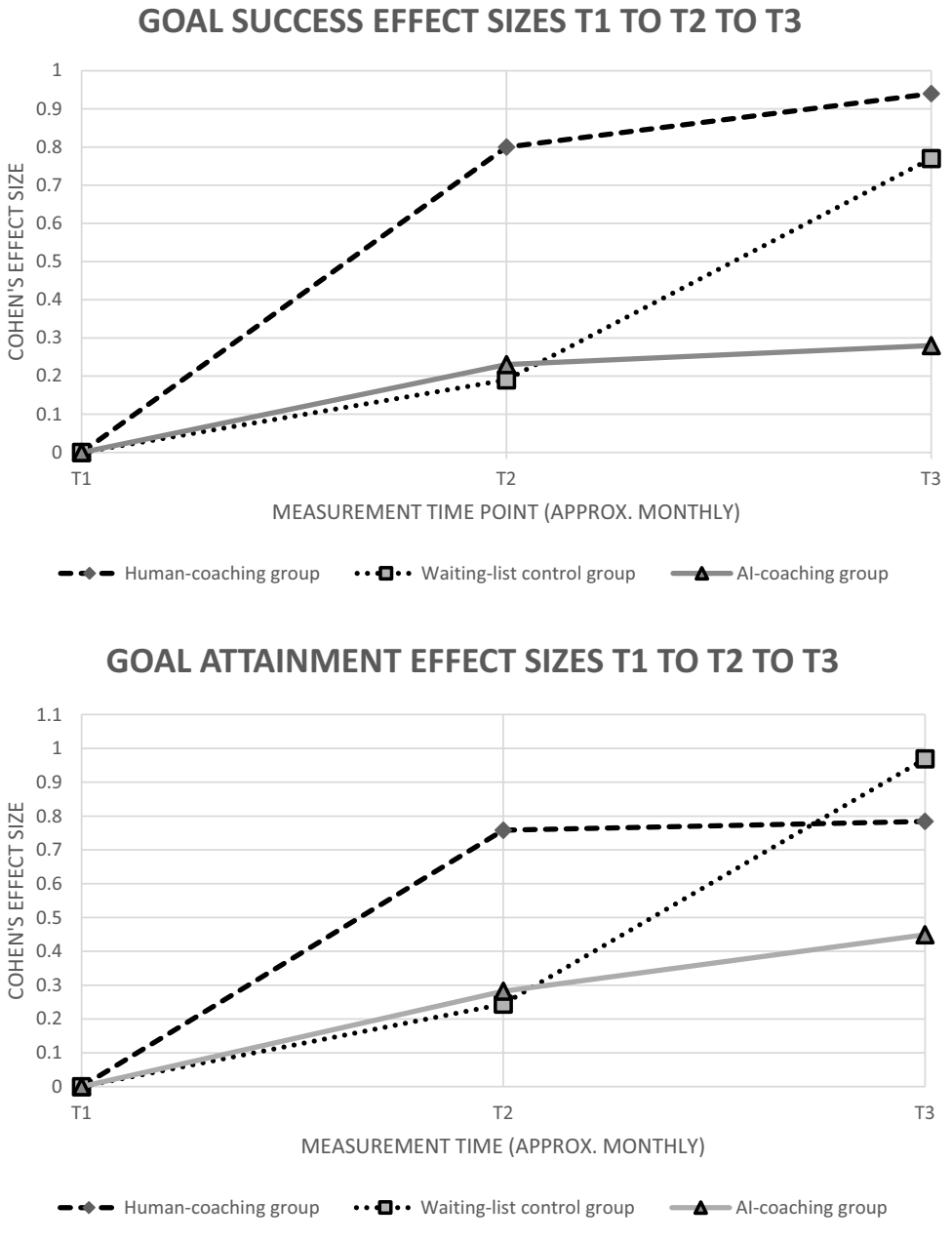

Figure 2. Cumulative effect size for (self-scored) goal 1 success between control and experimental groups (Cohen, 1988) d effect size changes. The second graph shows the same cumulative effect size d for goal attainment scores.

correlate strongly amongst themselves, we assume that they are in large part motivational or 'readiness' factors for coaching (Table 4). The presence of each of these factors can be considered as contributing to overall coaching success, whereas their absence might thwart or diminish the impact of strong coach–coachee relationships and interactions on specific outcomes. The results of this study also support the use of these factors as

outcome measures used in coaching studies to demonstrate overall coaching success (De Haan, 2021).

First indications that RCT findings in coaching are not affected by attrition

Even if only as a first indication, we can report that our full sample behaved the same as the 13% reduced sample that we would get if we took all subjects out who did not make use of the coaching as was offered. This helps with the credibility of other randomised controlled trials (Burt & Talati, 2017; De Haan & Nilsson, 2023; Nicolau et al., 2023) where similar levels of attrition are commonplace. As our co-regulation model predicts, attrition in executive coaching might not be due to exhaustion, boredom, resistance, or negative outcomes for the coachee, but may look more like a conscious choice by mature leaders around value and time spent in coaching.

We have seen the attrition rate being around ten times higher in the AI-coaching cohort even though coachees had regular reminders and kept in relationship with the principal researcher. This attrition bias needs to be considered against the background that drop-out rates in clinical trials can be kept lower through frequent contact and feedback, whilst in the real world this is not practical and attrition rates are expected to be substantially higher (Nwosu et al., 2022)

Coachees continue to rate the coaching highly, but they seem to optimise their engagement with it, including its duration. It is well known that coaching effects tend to remain the same over a broad range of total session numbers (De Haan & Nilsson, 2023). A similar result for attrition was reported by Stefens et al. (2022) for two work-alliance measurements during coaching. It would be a remarkable result, because other clinical interventions, such as psychotropic and psychotherapeutic interventions, are much more dose- and attrition-dependent,[2] whilst outcome-study results do get seriously biased by the impact of attrition (for clinical apps and chatbots: Nwosu et al., 2022; for psychotropic agents: Leon et al., 2006). This seems to show that dropouts in psychiatry and psychotherapy occur more for clients who would have benefitted from staying the course of the treatment but lack intrinsic motivation or struggle to rationalise the intervention. In contrast, our findings indicate that coachees remained consistently motivated to participate in the sessions, voluntarily engaging as long as they perceived ongoing added value.

Limitations

Several limitations of the current study must be acknowledged. Firstly, there are reasons to expect biases in our study sample. Nearly all coachees were middle or senior managers in the same global corporation, and they were required to be in the possession of a smartwatch to collect momentary measurements of their cardiac rhythms. Although there was no underlying organisational demand for coaching that we know of, the fact that they had to volunteer might have helped to have a more naturalistic sample. As their line managers were also somewhat involved in the trial, some of the goal-attainment gains may have arisen from additional management support rather than the impact of the

coaching itself, but that would have been true for all three trial groups. However, since there were virtually no nested data and all participants as well as line managers working in a broad range of contexts and countries, there is little possibility of peers in any of the groups influencing their colleagues' scores.

Secondly, there is same-source bias expected in some of the reported effects (when coachee measures are compared). Even when different sources are employed (e.g. coaches' and line managers' data), they are obviously in a relationship and therefore not fully independent.

Thirdly, there are organisational biases that might limit generalisability: coachees were leaders in multinational education, so they had relatively high education levels, were rather cosmopolitan, working in the English language and mostly within a non-English-speaking country. Limitations regarding the generalisability of the study sample to managers in other roles, organisations, industries, or geographic regions should be acknowledged. While the inclusion of participants from over 30 countries and all continents may have helped mitigate some biases, it remains uncertain whether these findings can be broadly generalised.

Fourthly, the AI chatbot we tested made use of generative AI and cannot be compared directly with the chatbot tested previously by Terblanche et al. (2022). Participants received reminders for engagement with the chatbot, but sessions and attendance were not monitored in the same way as was automatically done in human coaching (by the coach). It is possible that a study design with a longer duration or different algorithms might have resulted in more significant findings. The fast-developing digital coaching landscape today makes it difficult to fully evaluate coaching outcome research that is available. We maintain that there is still considerable progress to be made in developing highly effective coaching chatbots, as they consistently fall short compared to human coaching. This is also evident in Terblanche et al. (2022), where chatbots had no measurable impact on coaching 'readiness' factors such as wellbeing. Additionally, both studies reported higher dropout rates and more complaints among chatbot users.

Implications and future directions

Implications for HRD and coaching theory

There are important theoretical implications of this research in combination with the other 37 RCT's that we now have in coaching (De Haan & Nilsson, 2023). Our most rigorous findings seem to show that a dyadic, *co-regulation* model is at work in the coaching relationship, where coach and coachee both conspire to maximise outcomes. This means that motivation and preparation are fuel for the coaching (similar to Freud's idea of positive transference at the beginning of the work) and that they are in turn being enhanced through human coaching, i.e. further deepened through coaching interventions and human collaboration. We have less evidence that this mutual strengthening of the relationship is also at work in AI-chatbot coaching, but this remains a focus for future coaching research. Adherence and dropout seem to be a problem for chatbots, as is often noted in clinical trials as well (e.g. Linardon & Fuller-Tyszkiewicz, 2020). More importantly, co-regulation is a subtle process using selected small cues and shared memories,

which the more superficial chatbots are likely incapable of emulating fully (as clearly argued by Smith et al., 2025).

Human-to-human, relational contact seems important. Anecdotally, some of the results were memorable: several coachees reached out to the main researcher to personally thank him for creating the opportunity and making the match. Moreover, quite a few of the human coaching relationships were extended after the first month of coaching and became paid for by the coachee's corporate organisation.

Implications for practice

These findings extend past work and may help guide the coaching industry and the development of evidence-based approaches to coaching, as well as inform us of the choices that are made in the recruitment, development, deployment and matching of executive coaches (Wycherley & Cox, 2008), including from an AI-ethics perspective (Diller, 2024). On a practical level, coaches can learn to enhance the coachee's motivation levels (including self-efficacy) and 'readiness' for coaching: their hardiness, support and mental wellbeing generally. Coaches can advise coachees in the initial contracting conversations to keep up their hope and motivation, and to look after well-being or resilience aspects as well as their goals. In our view, coaches should also keep inquiring into the quality of the working alliance, particularly from the perspective of their coachees, as it is another important predictor of overall effectiveness.

Within broader HRD interventions executive coaching may be the *only* intervention at our disposal which is customised (tailored) enough to develop the leader's personality, wellbeing and reputation in the workplace to make a real, meaningful difference including to entire leadership teams and organisations, at that relatively deep, personal level. We would warn organisations against elitism. Executive coaching is usually already for privileged and higher-ranking leaders. The predictive value of factors, such as hope, wellbeing, cognitive hardiness, social support and self-efficacy may encourage organisations to pre-select coachees who are not only of higher rank but also already possess these desirable qualities. They should never lose sight of the ones who score lower on these 'readiness' measures and who might have more of a need to be coached even if their predictions for successful outcomes may lag. Statistically they may still do better out of coaching as they are starting from a higher need or a more stressful starting position. In other words, one thing that we would not recommend based on these results is pre-selecting candidates for coaching based on motivational (e.g. self-efficacy) or 'readiness' (e.g. wellbeing) factors.

A qualitative study (Terblanche & Tau, 2024) involving a similar AI coaching chatbot found that junior employees valued the availability of the chatbot (their human mentors were mostly unavailable) and reported that it positively helped them in their career development through reflection and self-awareness. Perhaps, the niche for AI coaching is exactly this underserved population who have had fewer self-development opportunities with fewer resources available. This proposition is backed up by the more positive results of a large-scale RCT study that compared AI coaching to human coaching using a student population (Terblanche et al., 2022)

Future HRD research

Our RCT study deliberately included leaders rather than students and made use of several measures used in previously published coaching studies for external validity and generalisability. Workplace coaching was conducted under realistic conditions with managerial coachees. Future HRD research should replicate this design and address some of the limitations pointed out earlier, including increasing the length of the coaching programme, reinforcing and measuring the translation of goal attainment into long-term habit formation through a lengthy period of deliberate behavioural practice of new skills back on the job.

Future studies should also include head-to-head comparisons with other AI coaching chatbots, digital assistants and diverse platforms, to better understand individual and organisational factors that might augment traditional executive coaching programmes. Where possible, similar momentary measurements of cardiac rhythms should be included in the future research using ecologically valid measurements to help provide greater insight into the coaching relationship and process. The current study builds on the cautions and recommendations of Jack et al. (2017) regarding methodological challenges in adding such neurophysiological measures within coaching research designs. However, replication of such studies with a larger sample size and statistical power is recommended and would advance the coaching research literature.

Finally, future research with neurodiverse and cross-cultural inclusion of coachees and coaches is needed for the future (along with continued research on coaching skills, training and credentialing) to better understand the successful intersection in the coaching alliance that promotes success. Such future research will provide guidance about which executive coaching programmes offering both human and AI digital platforms are most suitable and recommended for specific individuals with diverse, atypical backgrounds. With the emerging growth of AI embedded in almost every aspect of HRD, career and professional development within organisations today, much is to be studied and learned to clarify and enhance the HRD literature (cf., De Haan, 2024).

Summary

We believe that this research can help potential buyers of coaching (HR, line managers and Learning & Development professionals) in at least four different ways. Firstly, executive coaching appears to be a significantly effective intervention even when the duration is short term. Secondly, the human-to-human relationship in executive coaching seems important to bring about the positive result of the engagement. Thirdly, coaching effectiveness can be further improved by coachees being mentally prepared, resilient and motivated, whilst coaching in turn improves their wellbeing and perceived social support, in a virtuous circle; and finally, co-regulation remains an important factor in effectiveness as it explains well how attrition does not artificially boost effectiveness in RCT studies. Future research should continue to explore head-to-head comparisons between conventional face-to-face interventions and other (AI) digital coaching

platforms and coachbots, as well as include physiological measures that are associated with coachee and coach–coachee interactions.

Digital coaching applications are in the early stages of implementation in attempts to save costs, democratise coaching for all employees and provide ongoing support to coaches over time. Our study provides an ecologically valid response as to whether AI digital coaching efforts can augment, replace, or be considered effective by some employees, with specific presenting issues and across individual differences in light of current concerns surrounding ethical considerations in the use of digital and AI coaching (e.g. Diller, 2024).

Notes

1. One subject in our sample did request a change in coach after the first session, which we honoured, as most coaching providers would do.
2. In this regard, it was interesting to see that in the meta-analysis study of smartphone interventions in mental health (Linardon & Fuller-Tyszkiewicz, 2020), dose-dependency was found in all psychotherapeutic interventions but not in the 'general mental health' interventions which are much more like coaching. Likewise, the 'general mental health' interventions suffered significantly less from attrition, as if the co-regulation model applies there as well.

Acknowledgements

We want to thank our statistician Dr Dawit Assefa at Ashridge profusely for his flexibility and conscientiousness in running all the tests. Author 1 has no conflict of interest to report. He played a lead role in conceptualization, visualization, writing–original draft, writing–review, editing and research. Author 2 is an AI developer and has a commercial interest in the AI chatbot used in this research. He played a lead role in conceptualization, visualization, writing–original draft, writing–review, editing and literature research. Author 3 is the author of the cognitive hardiness scale used in this research and has a commercial interest in the measure. He played a supporting role in conceptualization, writing, review and editing, and literature research involved in the article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Erik de Haan http://orcid.org/0000-0001-9135-1110
Nicky Terblanche http://orcid.org/0000-0002-6796-6008
Kenneth Nowack http://orcid.org/0000-0002-9400-252X

References

Anthony, A., Gimbert, B., & Fultz, D. (2013). The effect of e-coaching attendance on alternatively certified teachers' sense of self-efficacy. *Journal of Technology & Teacher Education*, *21*(3), 277–299.

Athanasopoulou, A., & Dopson, S. (2018). A systematic review of executive coaching outcomes: Is it the journey or the destination that matters the most? *Leadership Quarterly*, *29*(1), 70–88. https://doi.org/10.1016/j.leaqua.2017.11.004

Barger, A. S. (2025). Artificial intelligence vs. human coaches: Examining the development of working alliance in a single session. *Frontiers in Psychology*, *15*, 1364054. https://doi.org/10.3389/fpsyg.2024.1364054

Bridgeman, J., & Giraldez-Hayes, A. (2024). Using artificial intelligence-enhanced video feedback for reflective practice in coach development: Benefits and potential drawbacks. *Coaching: An International Journal of Theory, Research & Practice*, *17*(1), 32–49. https://doi.org/10.1080/17521882.2023.2228416

Burt, D., & Talati, Z. (2017). The unsolved value of executive coaching: A meta-analysis of outcomes using randomised control trial studies. *International Journal of Evidence Based Coaching & Mentoring*, *15*(2), 17–24. https://doi.org/10.24384/000248

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health & Social Behavior*, *24*, 385–396. https://doi.org/10.2307/2136404

Crutzen, R., Viechtbauer, W., Kotz, D., & Spigt, M. (2013). No differential attrition was found in randomized controlled trials published in general medical journals: A meta-analysis. *Journal of Clinical Epidemiology*, *66*(9), 948–954. https://doi.org/10.1016/j.jclinepi.2013.03.019

Cuijpers, P., Karyotaki, E., De Wit, L., & Ebert, D. D. (2020). The effects of fifteen evidence-supported therapies for adult depression: A meta-analytic review. *Psychotherapy Research*, *30*(3), 279–293. https://doi.org/10.1080/10503307.2019.1649732

Dahling, J. J., Taylor, S. T., Chau, S. L., & Dwight, S. A. (2016). Does coaching matter? A multilevel model linking managerial coaching skill and frequency to sales goal attainment. *Personnel Psychology*, *69*(4), 863–894. https://doi.org/10.1111/peps.12123

De Haan, E. (2021). *What works in executive coaching: Understanding outcomes through quantitative research and practice-based evidence*. Routledge.

De Haan, E. (2024). A unique relationship: What role for AI in coaching? *Coaching Today Technology*, 14–18.

De Haan, E., Duckworth, A., Birch, D., & Jones, C. (2013). Executive coaching outcome research: The contribution of common factors such as relationship, personality match, and self-efficacy. *Consulting Psychology Journal: Practice and Research*, *65*(1), 40–57. https://doi.org/10.1037/a0031635

De Haan, E., Grant, A., Burger, Y., & Eriksson, P.-O. (2016). A large-scale study of executive coaching outcome: The relative contributions of working relationship, personality match, and self-efficacy. *Consulting Psychology Journal: Practice and Research*, *68*(3), 189–207. https://doi.org/10.1037/cpb0000058

De Haan, E., Gray, D. E., & Bonneywell, S. (2019). Executive coaching outcome research in a field setting: A near-randomized controlled trial study in a global healthcare corporation. *Academy of Management Learning and Education*, *18*(4), 1–25. https://doi.org/10.5465/amle.2018.0158

De Haan, E., Molyn, J., & Nilsson, V. O. (2020). New findings on the effectiveness of the coaching relationship: Time to think differently about active ingredients? *Consulting Psychology Journal*, *72*(3), 155–167. https://doi.org/10.1037/cpb0000175

De Haan, E., & Nilsson, V. O. (2023). What can we know about the effectiveness of coaching? A meta-analysis based only on randomized controlled trials. *Academy of Management Learning and Education*, *22*(4), 641–661. https://doi.org/10.5465/amle.2022.0107

Diller, S. J. (2024). Ethics in digital and AI coaching. *Human Resource Development International*, *27*(4), 584–596. https://doi.org/10.1080/13678868.2024.2315928

Dixon, L. J., & Linardon, J. (2020). A systematic review and meta-analysis of dropout rates from dialectical behaviour therapy in randomized controlled trials. *Cognitive Behaviour Therapy*, *49*(3), 181–196. https://doi.org/10.1080/16506073.2019.1620324

Dyrbye, L. N., Shanafelt, T. D., Gill, P. R., Satele, D. V., & West, C. P. (2019). Effect of a professional coaching intervention on the well-being and distress of physicians: A pilot

randomized clinical trial. *Journal of the American Medical Association Internal Medicine*, *179* (10), 1406–1414. https://doi.org/10.1001/jamainternmed.2019.2425

Egan, T., & Song, Z. (2005). *A longitudinal quasi-experiment on the impact of executive coaching* [Paper presentation]. 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Ekuma, K. (2024). Artificial intelligence and automation in human resource development: A systematic review. *Human Resource Development Review*, *23*(2), 199–229. https://doi.org/10.1177/15344843231224009

Ellinger, A. D., & Kim, S. (2014). Coaching and human resource development: Examining relevant theories, coaching genres, and scales to advance research and practice. *Advances in Developing Human Resources*, *16*(2), 127–138. https://doi.org/10.1177/1523422313520472

Fontes, A., & Dello Russo, S. (2021). An experimental field study on the effects of coaching: The mediating role of psychological capital. *Applied Psychology*, *70*(2), 459–488. https://doi.org/10.1111/apps.12260

Goff, P., Goldring, E., Guthrie, J., & Bickman, L. (2014). Changing principals' leadership through feedback and coaching. *Journal of Educational Administration*, *52*(5), 682–704.

Grant, A. M. (2008). Personal life coaching for coaches in training enhances goal attainment, insight, and learning. *Coaching: An International Journal of Theory, Research & Practice*, *1*(1), 54–70. https://doi.org/10.1080/17521880701878141

Grant, A. M. (2014). The efficacy of executive coaching in times of organisational change. *Journal of Change Management*, *14*(2), 258–280. https://doi.org/10.1080/14697017.2013.805159

Graßmann, C., & Schermuly, C. C. (2021). Coaching with artificial intelligence: Concepts and capabilities. *Human Resource Development Review*, *20*(1), 106–126. https://doi.org/10.1177/1534484320982891

Graßmann, C., Schölmerich, F., & Schermuly, C. C. (2019). The relationship between working alliance and client outcomes in coaching: A meta-analysis. *Human Relations*, *73*(1), 35–58. https://doi.org/10.1177/0018726718819725

Hamlin, R. G., Ellinger, A. D., & Beattie, R. S. (2008). The emergent 'coaching industry': A wake-up call for HRD professionals. *Human Resource Development International*, *11*(3), 287–305. https://doi.org/10.1080/13678860802102534

Hewitt, C. E., Kumaravel, B., Dumville, J. C., Torgerson, D. J., & Trial Attrition Study Group. (2010). Assessing the impact of attrition in randomized controlled trials. *Journal of Clinical Epidemiology*, *63*(11), 1264–1270. https://doi.org/10.1016/j.jclinepi.2010.01.010

Horney, K. (1936). The problem of the negative therapeutic reaction. *Psychoanalytic Quarterly*, *5* (1), 29–44. https://doi.org/10.1080/21674086.1936.11925271

Horvath, A. O., & Greenberg, L. S. (1986). The development of the working alliance inventory: A research handbook. In L. S. Greenberg & W. Pinsoff (Eds.), *Psychotherapeutic processes: A research handbook* (pp. 529–556). Guilford Press.

International Coaching Federation. (2023). 2023 ICF global coaching study. https://coachingfederation.org/resources/research/global-coaching-study/

Jack, A. I., Rochford, K., Friedman, P. M. A., Boyatzis, R., & Boyatzis, R. E. (2017). Pitfalls in organizational neuroscience: A critical review and suggestions for future research. *Organizational Research Methods*, *22*(1), 109442811770885. https://doi.org/10.1177/1094428117708857

Jones, R., Woods, S., & Guillaume, Y. (2015). The effectiveness of workplace coaching: A meta-analysis of learning and performance outcomes from coaching. *Journal of Occupational & Organizational Psychology*, *89*(2), 249–277. https://doi.org/10.1111/joop.12119

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: Generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, *18*(4), 1189–1220. https://doi.org/10.1007/s11846-023-00696-z

Kuntz, J. C., & Livingston, H. (2020). Surveying women leaders' career trajectories: Implications for leadership development in New Zealand organisations. *New Zealand Journal of Psychology*, *49*(2), 34–45.

Leon, A. C., Mallinckrodt, C. H., Chuang-Stein, C., Archibald, D. G., Archer, G. E., & Chartier, K. (2006). Attrition in randomized controlled clinical trials: Methodological issues in psychopharmacology. *Biological Psychiatry*, *59*(11), 1001–1005. https://doi.org/10.1016/j.biopsych.2005.10.020

Lin, L., Narender, R., & Zak, P. J. (2022). Why people keep watching: Neurophysiologic immersion during video consumption increases viewing time and influences behavior. *Frontiers of Behavioral Neuroscience*, *16*, 1053053. https://doi.org/10.3389/fnbeh.2022.1053053

Linardon, J., & Fuller-Tyszkiewicz, M. (2020). Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. *Journal of Consulting & Clinical Psychology*, *88*(1), 1. https://doi.org/10.1037/ccp0000459

Liu, X., & Batt, R. (2010). How supervisors influence performance: A multilevel study of coaching and group management in technology-mediated services. *Personnel Psychology*, *63*, 265–298. https://doi.org/10.1111/j.1744-6570.2010.01170.x

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *The American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Locke, E. A., & Latham, G. P. (2019). The development of goal setting theory: A half century retrospective. *Motivation Science*, *5*(2), 93.

Mai, V., Neef, C., & Richert, A. (2022). "Clicking vs. writing"-the impact of a chatbot's interaction method on the working alliance in AI-based coaching. *Coaching| Theorie & Praxis*, *8*(1), 15–31.

McGonagle, A. K., Beatty, J. E., & Joffe, R. (2014). Coaching for workers with chronic illness: Evaluating an intervention. *Journal of Occupational Health Psychology*, *19*(3), 385–398. https://doi.org/10.1037/a0036601

McGonagle, A. K., Schwab, L., Yahanda, N., Duskey, H., Gertz, N., Prior, L., Roy, M., & Kriegel, G. (2020). Coaching for primary care physician well-being: A randomized trial and follow-up analysis. *Journal of Occupational Health Psychology*, *25*(5), 297–314. https://doi.org/10.1037/ocp0000180

McLean, G. N., & González Ortiz de Zárate, A. (2024). Revolutionizing HRD through digitalization. *Human Resource Development International*, *27*(5), 756–775. https://doi.org/10.1080/13678868.2024.2399492

Merritt, S. H., Krouse, M., Alogaily, R. S., & Zak, P. J. (2022). Continuous neurophysiologic data accurately predict mood and energy in the elderly. *Brain Science*, *12*(9), 1240. https://doi.org/10.3390/brainsci12091240

Merritt, S. H., & Zak, P. J. (2024). Continuous remote monitoring of neurophysiologic immersion accurately predicts mood. *Frontiers in Digital Health*, *2*(6), 1397557. https://doi.org/10.3389/fdgth.2024.1397557

Millán, T. M., & Macours, K. (2017). *Attrition in randomized control trials: Using tracking information to correct bias*. SSRN.

Nicolau, A., Candel, O. S., Constantin, T., & Kleingeld, A. (2023). The effects of executive coaching on behaviors, attitudes, and personal characteristics: A meta-analysis of randomized control trial studies. *Frontiers of Psychology*, *14*, 1089797. https://doi.org/10.3389/fpsyg.2023.1089797

Nowack, K. M. (1990). Initial development of an inventory to assess stress and health risk. *American Journal of Health Promotion*, *4*(3), 173–180. https://doi.org/10.4278/0890-1171-4.3.173

Nwosu, A., Boardman, S., Husain, M. M., & Doraiswamy, P. M. (2022). Digital therapeutics for mental health: Is attrition the achilles heel? *Frontiers in Psychiatry*, *13*, 900615. https://doi.org/10.3389/fpsyt.2022.900615

Page, N., & De Haan, E. (2014). Does coaching work? . . . and if so, how? *Psychologist*, *27*(8), 582–586.

Passmore, J., Olafsson, B., & Tee, D. (2025). A systematic literature review of artificial intelligence (AI) in coaching: Insights for future research and product development. *Journal of Work-Applied Management*. https://doi.org/10.1108/JWAM-11-2024-0164

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. Measures in health psychology: A user's portfolio. *Causal and Control Beliefs*, *1*, 35–37.

Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, *20(6)*, 1081–1099.

Snyder, C. R., Irving, L. M., & Anderson, J. R. (1991). Hope and health. *Handbook of Social and Clinical Psychology: The Health Perspective*, *162*(1), 285–305.

Sonesh, S. C., Coultas, C. W., Lacerenza, C. N., Marlow, S. L., Benishek, L. E., & Salas, E. (2015). The power of coaching: A meta-analytic investigation. *Coaching: An International Journal of Theory, Research & Practice*, *8*(2), 73–95. https://doi.org/10.1080/17521882.2015.1071418

Stefens, M., Rondeel, E., Templin, J., Brode, D., de Waart, E., de Jong, R., ten Hoeve-Rozema, J., Waringa, A., Reijnders, J., Jacobs, N., & Lataster, J. (2022). Longitudinal measurement invariance of the working alliance inventory-short form across coaching sessions. *BMC Psychology*, *10*(1), 277. https://doi.org/10.1186/s40359-022-00968-5

Stewart, L. J., Palmer, S., Wilkin, H., & Kerrin, M. (2008). The influence of character: Does personality impact coaching success? *International Journal of Evidence Based Coaching & Mentoring*, *6*(1), 32–42.

Tavis, A., & Woodward, W. (2024). *The digital coaching revolution: How to support employee development with coaching*. Technical Kogan Page.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., & Stewart-Brown, S. (2007). The warwick-edinburgh mental wellbeing scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes*, *5*(1), 63. https://doi.org/10.1186/1477-7525-5-63

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

Terblanche, N. H. D., & Tau, T. (2024). Exploring the use of a goal-attainment, artificial intelligence (AI) chatbot coach to support first-time graduate employees. *Industry and Higher Education*, *39*(3), 09504222241287090. https://doi.org/10.1177/09504222241287090

Terblanche, N. H. D., Wallis, G. P., & Kidd, M. (2023). Talk or text? The role of communication modalities in the adoption of a non-directive, goal-attainment coaching chatbot. *Interacting With Computers*, *35*(4), 511–518. https://doi.org/10.1093/iwc/iwad039

Terblanche, N., & Kidd, M. (2022). Adoption factors and moderating effects of age and gender that influence the intention to use a non-directive reflective coaching chatbot. *SAGE Open*, *12*(2), 21582440221096136. https://doi.org/10.1177/21582440221096136

Terblanche, N., Molyn, J., De Haan, E., & Nilsson, V. O. (2022). Comparing artificial intelligence and human coaching goal attainment efficacy. *PLOS ONE*, *17*(6), e0270255. https://doi.org/10.1371/journal.pone.0270255

Terblanche, N., & Prywes, Y. (2025). An exploration of the role of visuals and users' imagery and verbal preferences on goal attainment and coaching chatbot adoption. *International Journal of Evidence Based Coaching & Mentoring*, *23*(1), 205–221.

Theeboom, T., Beersma, B., & van Vianen, A. E. M. (2014). Does coaching work? A metaanalysis on the effects of coaching on individual level outcomes in an organizational context. *The Journal of Positive Psychology*, *9*(1), 1–18. https://doi.org/10.1080/17439760.2013.837499

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the working alliance inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*, 207–210. https://doi.org/10.1037/1040-3590.1.3.207

Tschacher, W., Rees, G. M., & Ramseyer, F. (2014). Nonverbal synchrony and affect in dyadic interactions. *Frontiers in Psychology*, *5*, 1323. https://doi.org/10.3389/fpsyg.2014.01323

Ungerer, C., Heinzelmann, N., Baltes, G. H., & König, M. (2019). The effect of business coaching on NTBF survival-findings and lessons learned from a randomized controlled trial. In *2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (pp. 1–10). IEEE.

Wang, Q., Lai, Y.-L., Xu, X., & McDowall, A. (2021). The effectiveness of workplace coaching: A meta-analysis of contemporary psychologically informed coaching approaches. *Journal of Work-Applied Management*, *14*(1), 77–101. https://doi.org/10.1108/JWAM-04-2021-0030

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417–430. https://doi.org/10.1016/j.chb.2017.06.032

Wycherley, I. M., & Cox, E. (2008). Factors in the selection and matching of executive coaches in organisations. *Coaching: An International Journal of Theory, Research & Practice*, *1*, 39–53. https://doi.org/10.1080/17521880701878075

Zak, P. J., & Barraza, J. A. (2018). Measuring immersion in experiences with biosensors. In A. Cliquet Jr, S. Wiebe, P. Anderson, G. Saggio, R. Zwiggelaar, H. Gamboa, A. Fred, & S. B. i Badia (Eds.), *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies* (Vol. 4., pp. 303–307). https://doi.org/10.5220/0006758203030307

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. *Journal of Personality Assessment*, *52*(1), 30–41. https://doi.org/10.1207/s15327752jpa5201_2

Zimmermann, L. C., & Antoni, C. H. (2020). Activating clients' resources influences coaching satisfaction via occupational self-efficacy and satisfaction of needs. *Zeitschrift für Arbeits-und Organizationspsychologie*, *64*(2), 149–169. https://doi.org/10.1026/0932-4089/a000327