Preprint version, 3 May 2016

Evidence of competency:

Exploring coach, coachee and expert evaluations of coaching

James Lawley and Susie Linder-Pelz

Published 1 June 2016 online in

Coaching: An International Journal of Theory, Research & Practice

Available at http://dx.doi.org/10.1080/17521882.2016.1186706

James Lawley is an independent researcher and partner in The Developing Company which provides training and consultancy in the areas of modelling, metaphor and Clean Language. He is co-author of *Metaphors in Mind: Transformation through Symbolic Modelling* and has been a coach and supervising psychotherapist registered with the United Kingdom Council for Psychotherapy since 1993.
independent.academia.edu/JLawley and cleanlanguage.co.uk

Susie Linder-Pelz is an independent researcher with experience in academia as well as coaching. She gained a PhD from Columbia University and worked in behavioural science research before establishing a career coaching practice. Susie has authored 27 articles in peer-reviewed journals, 19 reports of commissioned research and five books including *NLP Coaching: An Evidence-Based Approach* (Kogan Page, 2010).
independent.academia.edu/SusieLinderPelz

Abstract

Competency-based coach training and assessment implies that coaching skills and effectiveness are closely related. But who is best placed to determine 'effectiveness'? While there are some studies comparing coach and coachee evaluations of coaching, none compare a coachee's evaluation with a coach trainer-assessor's rating of the coach's competency in the same encounter. Neither are there studies using coach, coachee and assessor triads.

This paper reports on research that examined how closely the evaluations of coachees, expert-assessors and coaches correspond. The research used a novel multi-method approach to triangulation including Clean Language interviewing (CLI) to explore coachees' experience and evaluation of coaching.

Assessor and coachee evaluations of the same coaching session were often at variance, both in terms of descriptive evaluations and numerical ratings. This suggests that compliance — or not — to a coaching methodology does not necessarily guarantee coachee satisfaction. While coach and coachee ratings showed no clear differences, in every triad coaches rated their own coaching considerably better than did the assessor.

Practical implications include the need for multiple sources of evidence to establish coach effectiveness and certification standards, the need for coaches to develop calibration skills so they can be more responsive to the coachees' in-session evaluations, and the usefulness of CLI together with established tools in evaluation research.

Keywords: coach competency assessment, evaluation research, coach effectiveness, Clean Language, Meta-Coaching, triangulation, calibration.

Practice Points

Relevant fields of practice are: the assessment and certification of coaches, and mixed-methods coaching research.

Primary contribution: The finding that coaches' compliance — or not — to a coaching methodology does not necessarily guarantee coachee satisfaction suggests the need for multiple sources of evidence, including outcome measures, to establish coach competency and standards.

Secondary contribution: An innovative prospective research design involving triads of coach, coachee and expert-assessor.

Tangible implications:

- Assessment of coach competency requires evaluation of both skills and outcome;
- Coaches can benefit from developing calibration skills of coachees' in-session evaluations;
- Clean Language, together with established evaluation tools, is useful in evaluation research;
- Assessment of coaching can be sensitive to the timing, method and sources of evidence used; further research is needed into how coaches, coachees and experts make assessments and how all three can contribute to a multi-perspective approach.

Introduction

Coach competency assessment

Competencies have emerged as crucial tools for appraisals and setting standards (Grant, Passmore, Cavanagh & Parker, 2010). Research to identify the most influential competencies continues to grow. A recent systematic review found 32 studies which investigated the coach attributes involved in a constructive coaching relationship or 'alliance' (Lai & McDowall, 2014); many consider the coaching alliance is key to coach effectiveness (Greif, 2007; O'Broin & Palmer, 2010).

Competency-based assessments rest on the principle that evaluation of observed behaviour will predict future performance (Potgieter & Van der Merwe, 2002). Although many accreditation bodies use competency-based assessments, it is unclear to what extent these are based on data and analysis (Lai, 2014). Similarly, there is little research into 'real-life' assessment of coaches.

Rationale for competencies

Competency-based frameworks have been constructed using a variety of sources: literature reviews; surveys and interviews of trainers, coaches and coachees; direct observations of coaching dyads; and benchmarking exemplary coach behaviours among others (Lai & McDowall, 2014). Competency research seeks to determine the degree to which various aspects of the coaching, coach or coachee affect outcome (de Haan & Duckworth, 2013). While results show there are observable 'success factors' or 'active ingredients' which can to varying degrees predict beneficial outcomes, it is not yet possible to specify those outcomes (Greif 2007; de Haan, Duckworth, Birch & Jones, 2013). Even in medicine, research documenting the effects of outcome- and competency-based education and assessment is rare (Morcke, Dornan & Eika, 2013).

What is competency?

'Competency' is a characteristic of an individual that is causally related to effective performance (Boyatzis, 1982). More simply it is the ability to do something successfully, to consistently get a beneficial outcome. Hidden in these definitions are two assumptions which are central to coach assessment and certification. First, since 'ability' is an abstract concept it cannot be observed directly; it has to be inferred from repeated observations of actual behaviour (Potgieter & Van der Merwe, 2002). This requires someone to make a subjective assessment that the observed behaviour is sufficiently fit for purpose. *Who* makes the assessment, *when* and *where* they evaluate, *what* they evaluate for and *how* they do the evaluating will all influence the assessment. Second, the definition of 'competence' presupposes 'success' or 'effectiveness', which can only be assessed if there is a measure of beneficial outcome (Grant, 2014; Greif, 2007). Unlike a driving test where the outcome — safe driving — is well defined and can be assessed at the time, 'success' cannot be observed during a coaching session. The outcomes of coaching are often complex, difficult to define and not known until sometime after the event (Linder-Pelz & Lawley, 2015).

If assessing competency requires both an evaluation of behaviour and a measure of outcome, why is it that so many competency-based assessments and certifications do not include outcome measures? Instead, it is presupposed that an expert can evaluate a coach's competency solely by observing behaviour specified by the coaching method. Hence many assessments based on competency frameworks could more accurately be called tests of compliance with model or adherence to method and are in part a convenient substitute for measuring effectiveness.

Why outcome measures are not included in assessments

Several factors contribute to outcomes not being included in coach assessments. These include the lack of consensus about appropriate outcome measures (Greif, 2013; de Haan et al., 2013) as well as the time-consuming and challenging nature of outcome assessment (Grant et al., 2010). Furthermore, the simplest and most frequently used outcome measure, coachee evaluation (de Haan & Duckworth, 2013), is viewed with widespread scepticism. Coachee satisfaction ratings are often considered reaction-level data that do not necessarily indicate concrete results but are more reliable when collected longitudinally (Greif, 2013). Since outcome effects can be overestimated by coachees when collected immediately after an intervention (de Haan et al., 2013) they are not to be taken at face value (Grant, 2014). Others go further: 'Client feedback is often inversely correlated to coaching quality — it is one of the least reliable measures we have' (Clutterbuck, 2013).

Potential consequences of behaviour-only assessment

While recognising that coachee evaluations give a subjective perspective on a coach's effectiveness, there are potential consequences of leaving the opinion of the primary beneficiary out of the assessment. If competency is assessed only by the behaviour of the coach — without reference to its effectiveness — an old medical saying may apply to coaching: 'the operation was successful but the patient succumbed'. Coaches need to skilfully and flexibly tailor their behaviour to the specific needs of the coachee (O'Broin & Palmer, 2009). This can present coaches with a dilemma when they are assessed based on adherence-to-method. If they adapt to the idiosyncrasies of the coachee they may not be compliant with the model and so put their accreditation at risk — even if the coachee evaluates the coaching to be instrumental in achieving a successful outcome.

Assessing coach competency from different perspectives

If coaching evaluation is best done from multiple perspectives (Greif, 2007, 2013) logically the same should apply to coach assessment. In competency-based assessment there are three perspectives — that of coach, coachee and expert-observer. To date little is know about how these perspectives compare; no studies address this directly though some cast an oblique light on it.

Dyad studies

Two dyad studies have analysed the effect of the coachee-coach relationship on outcomes (Boyce, Jackson & Neal, 2010; de Haan et al., 2013). Both studies showed dissonance between coach and coachee ratings and both concluded that the coach's estimate of the strength of the relationship did not correlate with the coaching outcomes nor with the relationship as measured by the coachee. These studies analysed results *across* a number of dyads. They did not compare responses *within* the same dyad. One study that compared within-dyad coach and coachee ratings of the coach's skill found a significant correlation between the coaches and coachees (Grant & Cavanagh, 2007). However, the ratings were not correlated against any outcome measure and the coachees were participants on coaching programmes rather than real-life coachees.

Two studies used direct observation of coaching sessions to look for coaching behaviours that can predict coachee-specified outcomes (Greif, 2010; Ianiro, Schermuly & Kauffeld, 2013). Other approaches have used interviews with both coach and coachee to compare experiences of 'critical moments' (de Haan, Bertie, Day & Sills, 2010) and of the coach-coachee relationship (Jowett, Kanakoglou & Passmore, 2012). However, none of these studies assessed the competency of the coach in the way an expert does when assessing competency for accreditation purposes.

Triads

Despite the strong case for triangulation (Johnson, Onwuegbuzie & Turner, 2007), few studies have used this method to reduce biases that might result from assessments relying on a single source. Apart from his own research, Seamons (2006) cites two other studies of organisational coaching which analysed triads of coach, coachee and coachee manager or colleagues. However, in these studies one member of the triad was not privy to what happened in the coaching session. We know of no research that has compared an expert's rating of a coach's competency with the coachee's evaluation of the coaching received, nor any studies that have compared an expert's assessment with the self-evaluation of the coach. And none, apart from Grant & Cavanagh (2007), have compared coach and coachee evaluations of the coach's competency.

Multiple methods and perspectives

There is a need for evaluation research: to embrace a range of investigative paradigms (Grant et al., 2010); to utilise data from multiple sources and to employ pragmatic assessment methods (Greif, 2013); and to provide rich data on the autogenic subjective experiences of those involved in coaching (Stober, Wildflower & Drake, 2006). A study that triangulated the assessment of coaches' competence from the perspectives of coach, coachee and expert would address the lack of research in this area. We knew of no single instrument that would adequately gather the evaluations of the three perspectives involved and thus multi-method research (MMR) would be appropriate (Johnson et al., 2007; Creswell, 2011). The research set out to investigate three questions:

(1) Do expert assessments of coaching competencies correspond with coachees' evaluation of the coaching they received?

(2) Do coaches' self-ratings of their coaching competencies correspond with coachees' ratings of the value received from the coaching?

(3) Do coaches' self-ratings correspond with an expert's assessment of their coaching competencies?

Method

This study was part of a larger one that used Clean Language interviewing (CLI) to distinguish between events, effects, evaluations and outcomes during and after coaching (Linder-Pelz & Lawley, 2015). It involved a multi-method approach to collecting data from coaches, coachees and an expert who evaluated the same coaching encounter. In addition to verbal reports, the research incorporated the 'legitimate and valuable uses of numbers even in purely qualitative research' (Maxwell, 2010, p. 476), analysing verbal and numeric data separately before merging the findings in the interpretation (Plano Clarke, Creswell, O'Neil Green & Shope, 2008).

Study design

To simulate a real-life assessment our triangulation procedure required: a number of triads involving the same coaching method which had a proven competency-assessment process; an assessor experienced in using the assessment instrument; an instrument by which the coaches qualified in the coaching method could self-assess; and a means for the coachees to evaluate the coaching which did not predefine the criteria they used to make their assessment. The last point suggests an inductive phenomenological approach (Gyllensten & Palmer, 2007; O'Broin & Palmer, 2010). While the coachee and expert descriptions would reveal similarities and differences in the evaluation of coach behaviour and its effects, they would not reveal to what degree each regarded the coach as competent. So the research needed to compare coachee and expert ratings on a scale. To complete the triangulation these ratings needed to be compared to

the coach self-ratings.

Meta-Coaching

Meta-Coaching, a goal-focussed method based explicitly on cognitive-behavioural psychology (Linder-Pelz & Hall, 2008), was chosen because a standard, benchmarked list of specific behavioural indicators of competency is routinely used to observe, assess and certify trainee coaches' readiness to practise (Hall, 2011). Meta-Coaching was also chosen because the seven core Meta-Coaching skills — supporting, listening, questioning, meta-questioning, giving feedback, receiving feedback, and inducing states — correspond to many of the common factors identified in an effective coach-coachee alliance (de Haan et al., 2013; Lai & McDowall, 2014). The validity of assessing the Meta-Coaching competencies using a benchmarking score sheet has been previously documented (Linder-Pelz, 2014).

Meta-Coach training and benchmarking is explicitly predicated on the hypothesis that the more often one sees the behaviours of competencies in a coach, the more likely a client will evaluate the coaching to be successful in facilitating the goals she or he has for coaching (Hall, 2010; Hall, personal communication, August 7, 2011). To test Hall's hypothesis we needed to involve both an expert who could assess coaches' competencies during 'real' coaching sessions (rather than in trainings where coachees are trainee coaches) and coachees who could evaluate the coaching they received.

Protocols

Protocols were developed for selecting and briefing coaches and coachees, for data collection and for analysis of the interviews (Lawley & Linder-Pelz, 2014).

Six volunteer coachees were randomly assigned to one of three practising coaches previously certified in the Meta-Coaching methodology. Each coachee received a single, 90-minute Meta-Coaching session (which was video-recorded) and two subsequent Clean Language interviews. All coaching sessions took place on the same day in an office setting that was independent of any of the coaches' work premises.

Clean Language, originated by counselling psychologist David Grove, was chosen for the in-depth interviews because it is a relatively content- and bias-free method (Tosey, Lawley & Meese, 2014; van Helsdingen & Lawley, 2012). CLI adheres to a strict protocol that keeps the interviewer from introducing any content or leading questions into the conversation, ensuring that the descriptions and evaluations obtained are sourced exclusively from the interviewee's personal vocabulary and experience. It is unique in having a detailed method for assessing the 'cleanness' of an interview and therefore the authenticity of the data collected.

Consent from coachees and coaches was obtained and anonymity assured. Neither the expert nor the interviewer had access to each other's data or findings and coachees were informed that their coach would not be privy to anything said in the interviews. Each data source was analysed independently before being compared within triads as well as across triads in relation to the three research questions.

Selecting participants

Three coaches, six coachees and one expert-assessor were recruited.

Coaches. Responding to an invitation to the Meta-Coaching community, three accredited Meta-Coaches volunteered. All were women aged in their 30s and 40s, running their own businesses with paying clients. Each was randomly allocated two volunteer coachees.

The expert. The primary developer of Meta-Coaching and its benchmarking assessment process was invited to participate. He had trained and certified hundreds of coaches worldwide, including the three in this study. His brief was to rate the coaching skills demonstrated in each video-recorded session according to the established benchmarking criteria; he was not told about the design or planned analysis of this study.

Coachees. Convenience and purposive sampling was done through a request from the second author to colleagues and acquaintances. Of the 11 replies received from prospective coachees, six met our criteria of: (1) no prior experience of Meta-Coaching; (2) something meaningful they wanted to change in their life; (3) not currently seeing a coach, psychologist or psychotherapist; (4) never having been diagnosed with a major psychological disturbance; and (5) not knowing the interviewer. Coachees were aged from mid-30s to early 60s, five were women and three had received other coaching. Each coachee received one 90-minute coaching session. The topics they chose to work with included health, building a business, confidence at work, self-worth, a relationship concern and managing money.

Data collection and preparation

Data were obtained from the coaches, the expert and the coachees. Table 1 summarises the sources, methods, sequence and timing of data collected.

Table 1. Data collection: sources, instruments and times

Data on	*Coaches' self-rating of their core coaching skills*	*Coaches' self-rating of their coaching of each coachee in this study*	*The level of core skills demonstrated in a video recording of each coaching session*	*Coachees' experience and evaluation of coaching, as reported in first interview*	*Coachees' experience & evaluation of coaching as reported in second interview*
Data code	**G1**	**G2**	**M1 & M2**	**T1**	**T2**
When data collected	Immediately prior to coaching sessions	Within 24 hours of coaching session	A few weeks after coaching session	One to two days after coaching session	12 to14 days after coaching session
Source of data	Coaches	Coaches	Expert observer	Coachees	Coachees
Data collection instruments	GCSQ Likert 12-item questionnaire 1-7 scale.	Modified GCSQ questionnaire completed for each coaching session.	Meta-Coaching benchmarking form 27 with numerical rating (M1) and written comments (M2).	Face to face interviews using Clean Language, recorded and transcribed.	Telephone interviews using Clean Language, recorded and transcribed; a single 1-10 scale for coachee satisfaction.
Number of cases	3	6	6	6	6

Expert data

Meta-Coach certification requires the coach to reach 2.5 on a scale of 0 to 3.5 on all seven core skills, which are assessed by reference to numerous behavioural sub-skills and summarised on the 'Expert Benchmarking Form 27' (Lawley & Linder-Pelz, 2014).

In our study, the expert used video recordings of the coaching sessions to rate the coaches' competencies. Only five of the seven criteria were applicable to one-off sessions, and the overall assessment was calculated by averaging the ratings assigned to these five criteria. The two criteria related to feedback were excluded. In addition to his ratings, the expert wrote comments about the coaches' strengths and developmental steps.

Coachee data

The coachees were interviewed twice by the first author. The first time was in person two days after the coaching in the same premises where the coaching sessions took place. These interviews lasted between 37 and 51 minutes. The second interview, two weeks later by telephone, took between 10 and 22 minutes. All interviews were audio-recorded and professionally transcribed.

The first interview started with a very open question such as 'How did the session go?' Coachees were not asked directly about coach competencies or 'core themes' (Passmore, 2010, p. 51). Instead, evaluations were explored when the coachee raised them. Coachees described what did and did not work well, and they assessed the coaching using their own self-defined criteria.

At the end of the second interview, the interviewer asked a 10-point Likert-type scale question developed specifically for this study based on the criteria the coachee had indicated were the most important: '*On a scale of 1 to 10, where 1 is no value whatsoever, and 10 is the highest value, where would you put what you got from the coaching in relation to* [coachees' words for what they valued most]?'

Coach data

Before the coaching sessions, each coach self-assessed her own level of competency using Grant and Cavanagh's (2007) Goal-focused Coaching Skills Questionnaire (GCSQ). The questionnaire items address five key competencies related to goal-focused coaching (of which Meta-Coaching is an example): outcomes of coaching, working alliance, solution-focus, goal setting, and managing process-accountability. The GCSQ consists of 12 statements against which coaches rate themselves on a Likert scale of 1 (very strongly disagree) to 7 (very strongly agree). One of the statements relating to feedback was not applicable to a single coaching session and was excluded from the analysis. The remaining 11 ratings were averaged.

Within 24 hours after each coaching session the coaches completed a modified GCSQ (Lawley & Linder-Pelz, 2014) in which the original statements were reworded to address a single session rather than coaching competency in general.

Analysis

To address the research questions we undertook the following comparisons (codes relate to those in Table 1):

(1) The expert's written comments for each coaching session (M2) were compared and contrasted with the coachee's statements in the interview two days after the session (T1). The expert's numerical rating (M1) was compared with the rating given by the coachee two weeks after their coaching session (T2).

(2) The coach's self-rating after each session (G2) was compared with their coachee's rating (T2).

(3) The coach's self-rating after each session (G2) was compared to the expert's rating of the coaching session (M1).

In addition to comparing findings within triads, the findings across triads were analysed by examining the relative rankings of the scores.

Reliability

Written protocols applied consistently enabled a robust exploration of the research questions. The two researchers undertook comparisons independently, followed by discussion of their varying interpretations. The researchers documented their regular discussions regarding methodological issues and decisions.

The research involved a field study that evaluated a real situation using two established measures — the benchmarking of core Meta-Coaching skills and the GCSQ. Since these instruments were not entirely compatible they were regarded as sources of qualitative data along with the more recently developed Clean Language interview method. CLI is well prescribed and replicable (Tosey et al., 2014) indicating its trustworthiness (Sousa, 2014). To check how closely the interviews adhered to the CLI protocol, a team of experienced Clean Language practitioners and researchers (not involved in this study) allocated one of four 'cleanness' ratings to every interviewer question or statement (Lawley & Linder-Pelz, 2014):

- *Classically clean*: drawn from the original Clean Language question set using only the interviewee's words.

- *Contextually clean*: introduces only 'neutral' words based on the context of the research or logic inherent in the interviewee's information.

- *Mildly leading*: introduces words with the potential to lead but with no discernible affect on the interviewee's answers.

- *Strongly leading*: introduces words (especially metaphors), presuppositions, frames or opinions that cast doubt on the authorship of interviewee answers.

The tabulated results were used to arrive at a summary assessment for each interview. The reviewers concluded that the interviews adhered substantially to the CLI protocol and were appropriate for the purpose of this research.

Findings

Descriptive evaluations

The expert's written comments were compared with each coachee's verbal report of the same session (at T1) and were allocated to one of three categories: compatible statements; incompatible statements; coachee value statements not mentioned by the expert. Verbatim examples of each category are given in Tables 2–4.

Table 2 shows that the expert and coachee descriptions were compatible in two main areas: (1) coach presence, acknowledging, listening; and (2) questions, depth of probing. The former has been linked to a constructive coaching relationship and the latter is common to most coaching interventions.

Table 2. Expert assessments compatible with coachees' comments

Dyad	*Expert's assessments*	*Coachees' comments*
A	Strengths: A very quiet presence. A quiet listening and acknowledging and being with the client.	A sense of her actually listening for what was going on underneath.
B	Strengths: Giving acknowledgments throughout … staying present with a very quiet state and presence.	I can't speak too highly of [coach]. For me she really epitomised the sort of person that I would be perfectly happy to go and visit if I felt I needed to.
C	Coach's strength is her quiet presence, her ability to be still and calm and hold the space for the client to think.	I admired her neutrality throughout the whole session and she has an incredible memory.
	Strengths: Awareness questions [and] some strength with meta-questions.	She just asked the right questions.
D	Questions not really grounded, [should be] 'What do you have to do to get this? Do you have a plan? Steps? Stages? Resources for creating this plan and the changes?'	I left wondering … 'How am I going to lock this in?' … I didn't walk away with tools … to lock it in at a cellular level.
	Strengths: Quiet presence that can wait – that can use silence.	I liked the way that she stayed very neutral … The tone … was really, like, soothing, calming, which made me respond the same way.
E	No demonstration of in-depth probing.	It was very difficult to get beyond certain places because the questioning was too open in style.
F	Strength: Asking exploration questions.	She was very good at … making me aware … clever and interesting way she made me aware of what I was doing.
	Strength: Hearing specific words.	I thought she was a terrific listener.

Table 3 shows some notable divergence of opinion between expert and coachees relating to the *degree* of confrontation, challenge, probe, direction and goal orientation. For example, in dyad B the expert said the coach did not once confront whereas the coachee's first words in the interview were 'I found it very confronting. She challenged me on just about everything I said'. Similarly in dyad A, the expert remarked that the coach needed to 'be more directive [with] in-depth probing' since there was 'not one clarity check'. However, the coachee thought 'we went from the surface issue to something more deep' and the questions were 'narrowing down … she helped me converge on stuff'. Some coachees expressed dissatisfaction with the pace of the session and level of rapport with the coach but the expert

did not mention these. This suggests the coachees' evaluations of the *effects* of the coaches' behaviour was either not noticed or mis-calibrated by the expert.

Table 3. Expert assessments *in*compatible with coachees' comments

Dyad	***Expert's assessments***	***Coachees' comments***
A	[Needs to] be more directive, engage more to make it a dialogue. Then can challenge, confront … in-depth probing. Not one clarity check in the entire 90 minutes!	We went from the surface issue to something more deep … It was from the coach actually saying, 'Shall we drill down on that?' … The questions that were beyond that were much more narrowing down … she helped me converge on stuff.
	[Coachee response not mentioned]	There was an element that ended up feeling a bit condescending … there's a vulnerability in admitting your fears. It needs to be held. … Did it affect the efficacy of the program? Probably.
B	Did not once confront or probe in depth.	I found it very confronting. She challenged me on just about everything I said.
	Did not provide enough leading and direction.	She was very good from my perspective in picking up on what I was saying and drawing more out of me. [She] 'pushed' in a positive way, through her questions, to force me to look at myself.
	[Pace not mentioned]	I felt also that it went at a very rapid pace.
C	Not being directive enough. [No] in-depth probing. Coach let far, far too much go by without doing the clarity checks and without challenging it.	She was very good at pulling on the information, making me think … Even though I guess we jumped around, I could actually see that they were all pieces to the puzzle.
	[Rapport not mentioned]	I couldn't read [if] she understood where I was coming from, not that I thought she was judging me but I did feel like there was a lack of rapport.
D	Client needs more direction.	[She] kept me really glued on the one path … pulling me back to the topic and highlighting certain words, that's where the magic was.
	Things not heard such as 'Allergic to savings'.	So when I did say that phrase 'allergic to savings' that's when [the coach] knew and I knew and my whole body lit up, beamed from the inside out.
E	Coach confronted coachee only once.	It was quite confronting. A lot of insights. It was very draining. I came away feeling as if I'd been run over. But also quite a sense of calm.
F	A strength was repeating [words] then using them for acknowledgments throughout.	I found the reflective listening technique a little bit overdone … too obvious and I found it really a bit off-putting, because I knew that as soon as I said something I was going to get it back again, and I got distracted by that.

Table 4 presents a selection of statements indicating the value the coachees received from the coaching. None of these were mentioned by the expert, which indicates that his assessment was based almost exclusively on one half of the competency equation, coach behaviour (adherence to method), while the other half (the outcome the coachee valued) was virtually ignored.

Table 4. Value perceived by coachees not referred to by the experts

Dyad A	Most important in terms of genuine out-take was the having the experience of imagining myself without, I can say, the hang-ups … Operating at my best in that environment.
Dyad B	I left the session having developed a strategy to be able to deal with that conflict, which was extremely beneficial.
Dyad C	I got a lot out of it … coming up with certain things that would make me feel confident that would then help me to put myself out there … Before it was a chore and now it was like 'OK, I'll wake up early and I'll do this'.
Dyad D	I've tried [to] change it like times a million. But I never touched on the emotional, mental, blueprint [before] … It was just such an amazing discovery and I couldn't have done it, or attempt to do it, on my own.
Dyad E	It was productive and provided some very important insights... insights that will help me achieve some clarity. Clarity is what I wanted.
Dyad F	I really became engaged with the session when [the coach] actually provided me with new material … I made some quite good connections with, why I've done things and why I haven't done things. And that was good. Good to talk through and come to that 'aha' moment.

Numerical evaluations

Coach self-ratings

To establish whether coaches regarded their sessions as reasonable examples of their competence or whether any were substantially below par the coaches' self-assessment of their general skill level *prior* to the sessions (which ranged from 5.3–6.3 on the GCSQ scale of 1–7) was compared with their assessment of their actual performance *after* each session. Five of the six sessions were rated equal to or above the general skill level (6.2–6.8); the one exception was a little below (4.9). This suggests the coaches rated their sessions as reasonable examples of their competency.

Expert ratings

The scale of the Meta-Coaching benchmarking system is 0–3.5 with competency set at 2.5 or above. In this study the expert's overall rating of sessions ranged from 1.0 –1.7 — all well below the level that warrants certification.

Coachee ratings

After two weeks all coachees rated the coaching between 7 and 10 out of 10. Two mentioned that they had changed their evaluation at some point. One would have rated the session 8 if the first 30 minutes were excluded but overall scored it 7. Another coachee reduced her rating from 10 at the time of the session to 8 two weeks later because 'I've slipped back a little bit'.

Comparison of perspectives

Once they had been converted into equivalent scores out of 10, coachees' numerical ratings could be compared to those of the expert and coach (Table 5).

Table 5. Ratings converted to scores out of 10 for six coaching sessions

Session	Expert rating	Coach rating	Coachee rating
A	4.9	9.7	7.0
B	4.6	9.2	8.0
C	4.3	9.4	8.5
D	4.0	8.8	8.0
E	4.0	9.4	10.0
F	2.9	7.0	8.5
	___	___	___
Average	4.1	8.9	8.3

The expert's ratings were consistently below those of the coaches and the coachees. The expert marked every one of the coaching sessions below 5 out of 10, whereas the coaches and coachees assessed every session at 7 out of 10 or higher. The variation between expert and coaches is all the more remarkable because, apart from one session, the coaches rated their performance as equal to or better than their usual competency.

A comparison of the coach-coachee ratings shows a different picture: in every case the coach's rating is closer to the coachee's than is the expert's.

The *ranking* of the three sets of ratings in Table 5 shows that at the extremities the expert's assessment of a session was almost the inverse of the coachee's assessment. Two of the expert's lowest ratings (sessions E and F) correspond to two of the highest scores given by the coachees; and the expert's two highest ratings (A and B) correspond with two of the lowest ratings of any coachee. The third comparison — expert and coach ratings — shows more consistency, with agreement on the highest and lowest ranked sessions.

Answering the research questions

Research question 1: Expert and coachees often differed. While the expert's written feedback often matched the verbal evaluations of the coachee in relation to the coaching relationship or alliance and the value of questions, they were often at variance in terms of confronting/ challenge, leading/directing and the pace/rapport. These discrepancies were mirrored in the expert's numerical rating of the coaches' skills, which were consistently below the coachees' ratings of the same coaching session. We found no support for Hall's hypothesis that the more the core coaching competencies are observed in coaches, the more likely their clients will evaluate the coaching to be successful.

Research question 2: Coach and coachee ratings showed no clear differences, suggesting coaches were more 'in tune' with their coachee's evaluations than was the expert.

Research question 3: In every case coaches rated their own coaching considerably better than did the expert, who assessed their performance below the level that would gain certification for a new coach.

Discussion

Variability of coaching method and environment was minimised by selecting coaches certified in the same coaching methodology with a well-documented benchmarking assessment process, and by having all the coaching sessions take place under the same conditions. Yet both the verbal and numerical evaluations varied considerably.

Explaining the variations

Three aspects of the study may explain some of the variation in the evaluations: when the data were collected, the criteria used in making evaluations and the methods of evaluating. These are further discussed below and in the context of the study's limitations.

Timing of data collection

The evaluations took place at different times. Most importantly, the coachees had more time than the coaches or the expert to reflect on the value of the session and, in particular, on the value of the outcomes they experienced in the following two weeks. Neither coach nor expert had access to this information.

Evaluation criteria

Comparing the criteria by which coachees evaluated their coaches with the GCSQ framework and the Meta-Coaching benchmarks shows some overlaps and some differences. All three evaluations involved criteria related to the coach having a supportive style, increased coachee clarity or self-awareness, and goal setting. Similar criteria are regarded as core coaching skills

(Gyllensten & Palmer, 2007; de Haan, Culpin & Curd, 2011; Grant, 2014).

Differences in criteria that may explain some of the variation include the benchmarking of 'meta-questions', the attention to action planning in the GCSQ and the coachees' sensitivity to the pace of the session.

The GCSQ, more than Meta-Coach benchmarking, gives weight to the coachee's experience. Half the GCSQ questions ask the coach to consider the coachee's experience, whereas the benchmarking criteria are primarily about observable coach behaviour. This may explain why coach ratings were closer to coachee ratings than were the expert's ratings.

Methods of evaluating

Another reason coachee and expert evaluations differed may be related to the method of evaluating. Kahneman's (2011) *peak-end rule* states that memory of an event is more influenced by a peak representative moment and how we feel at the end of the event than by an aggregation of moment-by-moment experiences at the time. In this study the expert assessed many behaviours throughout the session, whereas coachees may have based their assessment on a few peak or critical moments (de Haan & Nieß, 2012).

The *end* part of Kahneman's peak-end rule may also have contributed to the variation in ratings since coachees seemed to give greater weight to the end of the session than did the expert. Passmore (2010) found coachees evaluated takeaway tasks both positively and negatively. This matches the current study where every interviewee made unsolicited reference to the latter stages of the session, mostly in relation to post-session tasks (Linder-Pelz & Lawley, 2015). By contrast, the expert made little comment about the latter part of the sessions, presumably because tasking did not form part of the core Meta-Coaching competencies.

Assessment Implications

The finding that sometimes coachee and expert opinion can be diametrically opposed suggests that when either is the sole basis for assessing a coach's competency there is a risk that important factors are undervalued. This is illustrated in the case of one coachee in particular: the expert assessed the coach's performance as well below competency and yet two weeks after the session the coachee, who had experienced other types of coaching, said 'I think it was a goldmine … I couldn't even put a price on it … It was just such an amazing discovery and I couldn't have done it, or attempt to do it, on my own … I honestly feel it did serve the exact purpose that I wanted it to serve and I can tick that off my list. Mission accomplished, literally.' Clearly the coaching had a valuable outcome for the coachee — and that is surely a primary purpose of coaching. Even though the expert's feedback to the coach included 'strengths', most of the feedback focused on how she did not meet the benchmarks rather than on what she did that worked for her coachee.

Trainee coaches can be placed in a difficult position when they perceive a conflict between facilitating what the coachee needs and complying with the competencies laid down by the coaching methodology.

This study suggests that experts and coachees are in part using different criteria and methods to evaluate, and that both viewpoints need to be considered in order to reflect the dual aspects of competency: ability and outcome. Rather than a one-perspective-fits-all approach to coach assessment, a more balanced approach might include an assessment method that takes into account three perspectives:

- Two or more experts assessing whether the coach demonstrates adherence to the coaching methodology.

- The coach evidencing an ability to calibrate how closely his or her practice adheres to the methodology *and* the value of the coaching to the coachee.
- The coachee's evaluation of their experience of the coaching.

Future research could investigate whether coach assessors can predict the ratings coachees give to their session. A large variation between expert and coachee evaluations would suggest a need to consider whether the assessment of coaching competencies is balanced or whether the criteria and benchmarks need expanding.

Practice Implications

If coaching — and in particular the coaching alliance — is largely a coachee-led process, there is value in coaches being attuned to the coachee (de Haan et al., 2013) and adapting to changes in the coachee's responses that occur throughout the session (O'Broin & Palmer, 2010).

'Calibration' means using a set of explicit or implicit criteria to assess a coachee's internal state from his or her verbal and nonverbal behaviour (Linder-Pelz, 2010). This is a broader definition than 'affect calibration', one of the 'eight observable success factors' identified by Greif, Schmidt & Thamm (2010, p. 5). We hypothesise that the more coaches can calibrate the coachee's on-going assessment of the value of the coaching, the better they will be able to support the coaching relationship and adapt what they are doing to what works well for the coachee. If coaches cannot accurately calibrate when they are — and especially when they are *not* — being effective, this could become an ethical issue. It follows that coaches who can consistently predict their coachees' ratings are likely to be more adept at calibrating their coachee's experience.

Future research could explore the accuracy of coaches' calibration skills. An informal investigation (led by the first author) asked 10 coaches to rate the session from their own

viewpoint and to estimate the coachees' rating for the session. The results suggest that there might be considerable room for improvement in coaches' calibration skills, especially at the extremes where coachees give an unusually high or low rating to a session (Lawley & Tompkins, 2014).

In addition, coaches and coachees could separately watch a video of their coaching session and, at particular moments, be asked to recall how they were assessing the value of the coaching at those moments. CLI could be used to tease out how coaches calibrate their coachees.

Limitations

Although we selected participants and collected data in a way that yielded information relating to the research questions (Saunders & Rojon, 2014), there was limited randomisation of participants and the findings of a small exploratory study may not be representative of coach assessment generally. The robustness of the research could have been increased if a different expert had certified these coaches previously and if more than one expert was involved in the assessment process, as occurs in the certification of Meta-Coaches.

Further studies could use a similar triangulation process with different coaching methodologies, expert-assessors, Clean Language interviewers as well as a larger number and diversity of coaches, coachees and assessors.

To minimise researcher bias we used reflexive discussions and crosschecked the results. The external verification of the adherence to Clean Language interview protocols and the detailing of our methods of data collection and analysis supports the study's credibility, replicability and 'auditability' (Ryan-Nicholls & Will, 2009). Verifying the findings with the participants would have further authenticated the results (Seamons, 2006).

As indicated earlier, the diversity of evaluation criteria and methods directed attention to different aspects of the coaching process (coach behaviour and coachee experience) and this may have affected the findings. It is challenging to relate data sets that do not address exactly the same concepts (Plano Clarke et al., 2008) and when different methods and forms of data are used (Ryan-Nicholls & Will, 2009). However, this study made no attempt to identify causal relations; instead it only compared evaluations of the same event from multiple perspectives. A future study may aim to reduce variability by asking two or more experts to also use the GCSQ.

Studying an on-going process can itself influence that process (Brannick & Coghlan, 2007; de Haan & Nieß, 2012). The purpose of the CLIs was to facilitate coachees' reflection on their evaluation of the coaching. Several coaches remarked that the interviews also gave them a better understanding of what had occurred. Although this suggests the research influenced the interviewees, the rigorous validation of the interview questions did not reveal any evidence that the interviews biased the substance of the interviewees' evaluations.

Conclusion

By documenting the degree to which verbal and numerical assessments of a coaching session can vary, depending on *who* is making the judgement and *when*, this study contributes to the debate about *how* best to assess coach competency. It suggests caution is needed when relying solely on coach behaviour assessments as determinants of coaching efficacy, as is common in many coaching training programmes. A more comprehensive picture will be obtained when the perspectives of expert, coach and coachee are all taken into account.

This study also demonstrates how a novel and pragmatic approach to triangulation of data from coachees, coaches and a coach assessor worked in practice. It lays the groundwork for further research using a similar triangulation methodology and shows how the use of

established evaluation tools together with Clean Language interviews can produce a rich source of information about subjective and individual matters.

In offering suggestions for coach practitioners, trainers and assessors, we propose, in particular, that coaches and assessors develop skills to better calibrate the on-going evaluation of coaching by coachees. This would enable coaches to be more responsive to what is happening in the moment and assessors to assess more than simply compliance with method.

Acknowledgements

We thank Dr Michael Hall for assessing the coaching sessions, Dr Paul Tosey for commenting on the draft manuscript and the coachees and coaches for their willingness to participate.

References

Boyatzis, R. E. (1982). *The competent manager: A model for effective performance*. New York: John Wiley and Sons.

Boyce, L. A., Jackson, R. J., & Neal, L. J. (2010). Building successful leadership coaching relationships: Examining impact of matching criteria in a leadership coaching program. *Journal of Management Development, 29*(10), 914-93.

Brannick, T., & Coghlan, D. (2007). In defense of being "native": The case for insider academic research. *Organizational Research Methods*, *10*(1), 59-74.

Creswell, J. W. (2011). Controversies in mixed methods research. In N. K. Denzin & Y.S. Lincoln (eds), *The SAGE handbook of qualitative research* (Ch 15). pp. 269-284. Thousand Oaks: SAGE.

Clutterbuck, D., (2013). Comment on coach assessment. *Euro-Coach List*, 23 April. Retrieved from www.eurocoachlist.com.

de Haan, E., Bertie, C., Day, A. & Sills, C. (2010). Critical moments of clients and coaches: A direct-comparison study. *International Coaching Psychology Review*, *5*(2), 109-128.

de Haan, E., Culpin, V. & Curd, J. (2011). Executive coaching in practice: What determines helpfulness for clients of coaching? *Personnel Review, 40*(1), 24-44.

de Haan, E., & Duckworth, A. (2013). Signalling a new trend in executive coaching outcome research. *International Coaching Psychology Review, 8*(1), 6-19.

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 American 65*(1), 40-57.

de Haan, E., & Nieß, C. (2012). Critical moments in a coaching case study: Illustration of a process research model. *Consulting Psychology Journal: Practice and Research*, *64*(3), 198-224.

Grant, A. M. (2014). Autonomy support, relationship satisfaction and goal focus in the coach–coachee relationship: which best predicts coaching success? *Coaching: An International Journal of Theory, Research and Practice, 7*(1), 18-38.

Grant, A. M., & Cavanagh, M. J. (2007). The goal-focused coaching skills questionnaire: Preliminary findings. *Social Behavior and Personality: An International Journal, 35*(6), 751-760.

Grant, A. M., Passmore, J., Cavanagh, M. J., & Parker, H. M. (2010). The state of play in coaching today: A comprehensive review of the field. *International Review of Industrial and Organizational Psychology, 25*(1), 125-167.

Greif, S. (2007). Advances in research on coaching outcomes. *International Coaching Psychology Review, 2*(3), 222-249.

Greif, S. (2010). A new frontier for research and practice: Observation of coaching behaviour. *The Coaching Psychologist, 6*(2), 21-29.

Greif, S. (2013). Conducting organizational-based evaluations of coaching and mentoring programs. In J. Passmore, David B. Peterson & T. Freire (eds), *The Wiley-Blackwell handbook of the psychology of coaching and mentoring* (Ch. 23). pp. 445-470. Chichester: John Wiley & Sons, Ltd.

Greif, S., Schmidt, F., & Thamm, A. (2010). The rating of eight coaching success factors—Observation manual version 4. Retrieved from www.home.uni-osnabrueck.de/sgreif/downloads/Rating_of_Coaching_Success_Factors_Version4-May_2010.pdf.

Gyllensten, K., & Palmer, S. (2007). The coaching relationship: An interpretive phenomenological analysis. *International Coaching Psychology Review, 2*(2), 168-177.

Hall, L. M. (2010). The facilitation model. In L. M. Hall, *Coaching conversations* (3rd ed.). pp. 145-154. Grand Junction, CO: Neuro-Semantics Publications.

Hall, L. M. (2011). *Benchmarking intangibles: The art of measuring quality.* Grand Junction, CO: Neuro-Semantics Publications.

Ianiro, P. M., Schermuly, C. C., & Kauffeld, S. (2013). Why interpersonal dominance and affiliation matter: An interaction analysis of the coach-client relationship. *Coaching: An International Journal of Theory, Research and Practice, 6*(1), 25-46.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

Jowett, S., Kanakoglou, K., & Passmore, J. (2012). The application of the 3+1Cs relationship model in executive coaching. *Consulting Psychology Journal: Practice and Research, 64*(3), 183-197.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Straus and Giroux.

Lai, Y-L. (2014). *The development and validation of a coaching psychologist competency framework through job role analysis: Focusing on the effective coaching relationship*. Paper presented at the Division of Occupational Psychology (DOP) Annual Conference, Brighton. (Supplied by author.)

Lai, Y-L., & McDowall, A. (2014). A systematic review (SR) of coaching psychology: Focusing on the attributes of effective coaching psychologists. *International Coaching Psychology Review, 9*(2), 120-136.

Lawley, J., & Linder-Pelz, S. (2014). Research protocols. Retrieved from www.cleanlanguage.co.uk/articles/articles/350/.

Lawley, J., & Tompkins, P. (2014). Calibration and evaluation–three years on. Retrieved from www.cleanlanguage.co.uk/articles/articles/348/.

Linder-Pelz, S. (2010). *NLP coaching: An evidence-based approach for coaches, leaders and individuals.* London: Kogan Page Limited.

Linder-Pelz, S. (2014). Steps towards the benchmarking of coaches' skills. *International Journal of Evidence Based Coaching and Mentoring, 12*(1), 47-62.

Linder-Pelz, S., & Hall, L. M. (2008). Meta-Coaching: A methodology grounded in psychological theory. *International Journal of Evidence Based Coaching and Mentoring*, *6*(1), 43-56.

Linder-Pelz, S., & Lawley, J. (2015). Using Clean Language to explore the subjectivity of coachees' experience and outcomes. *International Coaching Psychology Review, 10*(2), 161-174.

Maxwell, J. A. (2010). Using numbers in qualitative research. *Qualitative Inquiry, 16*(6), 475 -482.

Morcke, A. M., Dornan, T., & Eika, B. (2013). Outcome (competency) based education: An exploration of its origins, theoretical basis, and empirical evidence. *Advances in Health Science Education, 18*, 851-863.

O'Broin, A., & Palmer, S. (2009). Co-creating an optimal coaching alliance: A cognitive behavioural coaching perspective. *International Coaching Psychology Review, 4*(2), 184-194.

O'Broin, A., & Palmer, S. (2010). Exploring key aspects in the formation of coaching relationships: Initial indicators from the perspective of the coachee and the coach. *Coaching: An International Journal of Theory, Research and Practice*, 3(2), 124-143.

Passmore, J. (2010). A grounded theory study of the coachee experience: The implications for training and practice in coaching psychology. *International Coaching Psychology Review, 5*(1), 4-62.

Plano Clarke, V. L., Creswell, J. W., O'Neil Green, D. & Shope, R. J. (2008). Mixing quantitative and qualitative approaches: An introduction to emergent mixed methods research. In S. N. Hesse-

Biber & P. Leavy (eds), *Handbook of emergent methods* (pp. 363-387). New York: The Guilford Press.

Potgieter, T. E., & Van der Merwe, R. P. (2002). Assessment in the workplace: A Competency-based approach. *South African Journal of Industrial Psychology, 28*(1), 60-66.

Ryan-Nicholls, K. & Will, C. I. (2009). Rigour in qualitative research: Mechanisms for control. *Nurse Researcher, 16*(3), 70-85.

Saunders, M. N. K., & Rojon, C. (2014). There's no madness in my method: Explaining how your research findings are built on firm foundations. *Coaching: An International Journal of Theory, Research and Practice*, 7(1), 74-83.

Seamons, B. L. (2006). *The most effective factors in executive coaching engagements according to the coach, the client, and the client's boss*. (Unpublished doctoral dissertation). Saybrook Graduate School and Research Center, California, (UMI Dissertations Publishing 3206219).

Sousa, D. (2014). Validation in qualitative research: General aspects and specificities of the descriptive phenomenological method. *Qualitative Research in Psychology*, *11*(2), 211-227.

Stober, D. R., Wildflower, L., & Drake, D. (2006). Evidence-based practice: A potential approach for effective coaching. *International Journal of Evidence Based Coaching and Mentoring*, *4*(1), 1-8.

Tosey, P., Lawley, J., & Meese, R. (2014). Eliciting metaphor through Clean Language: An innovation in qualitative research. *British Journal of Management, 25*(3), 629-646.

van Helsdingen, A., & Lawley, J. (2012). Modelling shared reality: Avoiding unintended influence in qualitative research. *Kwalon: Journal of the Netherlands Association for Qualitative Research*, *1*(3). Retrieved from www.cleanlanguage.co.uk/articles/articles/328/