Behavioural and Cognitive Psychotherapy (2019), **47**, 736–744
doi:10.1017/S1352465819000316

MAIN

Judging clinical competence using structured observation tools: A cautionary tale

Anthony D. Roth[1,*], Pamela Myles-Hooton[2] and Amanda Branson[2]

[1]Research Department of Clinical, Educational and Health Psychology, University College London, London, UK and [2]Charlie Waller Institute, University of Reading, Reading, UK
*Corresponding author. Email: a.roth@ucl.ac.uk

(Received 30 July 2018; revised 01 February 2019; accepted 19 February 2019; first published online 15 April 2019)

Abstract

Background: One method for appraising the competence with which psychological therapy is delivered is to use a structured assessment tool that rates audio or video recordings of therapist performance against a standard set of criteria.
Aims: The present study examines the inter-rater reliability of a well-established instrument (the Cognitive Therapy Scale – Revised) and a newly developed scale for assessing competence in CBT.
Method: Six experienced raters working independently and blind to each other's ratings rated 25 video recordings of therapy being undertaken by CBT therapists in training.
Results: Inter-rater reliability was found to be low on both instruments.
Conclusions: It is argued that the results represent a realistic appraisal of the accuracy of rating scales, and that the figures often cited for inter-rater reliability are unlikely to be generalizable outside the specific context in which they were achieved. The findings raise concerns about the use of these scales for making summative judgements of clinical competence in both educational and research contexts.

Keywords: assessment; clinical supervision; therapist competence; training

Introduction

There are many reasons for developing scales to assess the competence with which a psychological therapy is delivered. For example, researchers may need to establish whether therapists in a clinical trial are adherent to a particular method, and competent in its delivery. 'Adherence' and 'competence' are related but conceptually distinct terms, the former referring to whether a therapist uses strategies and techniques identified as relevant to an approach, and the latter to the level of skill shown when implementing these techniques (e.g. Waltz *et al.*, 1991). As such, competence is a matter of 'doing the right thing in the right way', making it possible to be adherent but not competent.

In a training context scales can be used to conduct summative assessments of trainee progression, or are used formatively as part of supervision. A number of scales have been developed to gauge adherence and competence in the delivery of cognitive behaviour therapy (CBT; Muse and McManus, 2013). Of these, the most extensively researched is the Cognitive Therapy Scale (CTS: Young and Beck, 1980) and its later revision (CTS-R: Blackburn *et al.*, 2001). These measures typically rely on the judgement of raters in the structured assessment of audio or video recordings of psychotherapy sessions, so inter-rater reliability and scale validity are key requirements.

Loades and Armstrong (2016) report a systematic review of 20 studies that have investigated the inter-rater reliability of the CTS and its variants. Some had a primary aim of investigating the metrics of the scale, but most employed the CTS in the service of a relevant research question (for

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

example, in studies relating therapist competence to outcome) and reported on inter-rater reliability as part of the study design. Of the 20 studies, nine reported on the use of the CTS or CTS-R applied to clinical work with adults with anxiety or depression, while the remainder reported on adaptations of the CTS intended to make it more applicable to specific client populations (such as people with psychosis, with social anxiety disorder or to children). Intra-class correlation coefficients (ICCs; a measure of the reliability of ratings[1]) varied widely across studies, from 0.40 to 0.98, with a median of 0.65. This variation is also seen within studies: McManus *et al.* (2012) report separate evaluations of recordings of trainees early and late in their training; ICCs for these two time-points were 0.47 and 0.71, respectively.

This wide variation in reliability estimates merits further exploration. Differences in estimates of inter-rater reliability may reflect factors such as the degree to which raters were trained and relatedly the extent to which raters have improved concordance by discussing any differences in their interpretation of the scale to achieve consensus. While it seems that groups of raters working together in this way can achieve very high levels of reliability (Loades and Armstrong, 2016), where they are working more independently there seems to be poorer agreement. For example, in Dimidjian *et al.* (2006) three raters appraised recordings; two were 'in-house' to the research team and one was an external expert. The overall ICC for the raters working together was 0.94, but this reduced to 0.47 with the inclusion of the 'external' rater, this despite the fact that all the raters were highly expert in CBT, both as trainers and developers. A similar (if more numerically extreme) picture is reported by Jacobson and Gortner (2000) where the ratings of two 'external' assessors (selected to be both expert and independent) were contrasted to each other and to an internal rater, yielding ICCs between 0.01 and 0.08.

It is clear (and not altogether surprising) that groups of raters can work towards a consensual position in which their ratings are closely calibrated, and so achieve good inter-rater reliability. However, *consistency* in ratings does not speak to the 'accuracy' of the judgements being made; reliability does not equate to validity. Scores from different groups of raters may be at a different level, within-group ratings being concordant, but between-group ratings being discrepant. This observation is particularly pertinent if the rating scale is being used to make summative assessments, for example appraising a trainee's competence to practise. At issue is the level of concordance achieved by raters who are working independently with no or minimal training or active coordination; in other words the reliability and validity of the instrument when used in the field. As such a key aim of this paper is to establish the extent to which measures are appropriately used in routine circumstances for formative and summative evaluation of competence.

A new scale has been developed for structured observation of CBT; it is fully described in Roth (2016). This scale is rooted in the competence framework for CBT[2] (Roth and Pilling, 2008), which was developed as part of the English Improving Access to Psychological Therapy (IAPT) programme; this framework was used to generate the IAPT CBT curriculum for working with people presenting with anxiety and depression (information about IAPT can be found at: https://www.england.nhs.uk/mental-health/adults/iapt/). The framework organizes the delivery of CBT into discrete areas of activity, and identifies the knowledge and skills that underpin all variants of CBT as well as specific CBT skills that are applied when working with specific conditions or presentations. A distinctive aspect of the UCL[3] CBT scale is its identification of intervention methods that are present in almost all sessions along with those which characterize evidence-based interventions for specific disorders.

The framework also includes a domain of Generic Therapeutic Competences, knowledge and skills that are common across therapy modalities (for example, relational competences such as

[1] Conventionally, values less than 0.5 are taken to indicate poor reliability, between 0.5 and 0.75 moderate reliability, between 0.75 and 0.9 good reliability, and values greater than 0.90 excellent reliability.

[2] Accessed at: www.ucl.ac.uk/core/competence-frameworks

[3] University College London.

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

alliance building and repair) and skills associated with the management of sessions (for example, using measures, responding to emotional expression, or ending sessions). Although generic competences are necessary skills for the effective delivery of therapy, it is helpful to separate them from CBT-specific skills; by definition they are non-specific, and so do not test how well a therapist is applying CBT. As such, two parallel scales were developed, both of which would usually be administered, focusing on generic and CBT competences, respectively. Unlike measures developed for research use, the UCL scales are intended to be used in routine service contexts without extensive training, on the basis that each item is anchored with descriptions of specific therapist behaviours.

The present study has two aims. First, to contrast the inter-rater reliability of the UCL scales against the CTS-R in a context where raters are effectively working independently (as would be the case in most routine settings) rather than as part of a team. Given that a plausible hypothesis is that levels of inter-rater reliability on all three scales will be low, a second aim is to consider any implications for the use of therapy competence measures for formative and summative evaluation of competence in routine settings.

Method

Ethics

Ethical approval for this study was obtained from the UCL Research Ethics Committee (CEHP/2013/507). All clients gave written informed consent for their recordings to be used as part of this research study. Clients whose recordings were included in the trial were informed that their recordings could be used for educational research at the same time as their consent was obtained for recording sessions for training purposes.

Setting

The study was conducted at a university offering training in cognitive behaviour therapy for people with depression and anxiety presentations (as part of the IAPT programme).

Therapists

Fourteen therapists contributed recordings to the study (eleven women and three men). Their mean age was 32 years (range 26–42); all were registered on a one-year Postgraduate Diploma offering training in CBT as part of the IAPT programme. Therapists were of different professional backgrounds, varied in relation to their experience of mental health presentations, and had varying levels of prior exposure to CBT (although all had at least two years of clinical experience, and seven some had experience of delivering self-help CBT programmes in their previous roles).

Trainees on this programme routinely submit session recordings for evaluation, and the sessions for this study were selected from this corpus. There was no attempt to select recordings systematically in relation to the therapists' prior experience, or their stage of training.

Clients and treatments

All clients were seen in the setting of the IAPT services in which their therapists were employed. All were adults aged between 19 and 62 years referred with a primary diagnosis of depression or with an anxiety disorder (phobia, panic disorder, generalized anxiety disorder or social anxiety). CBT interventions were mapped to presentation, guided by the IAPT training curriculum for CBT (www.babcp.com/files/Accreditation/Course/dh_083169.pdf); interventions for depression followed Beck's model (Beck *et al.*, 1979).

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

Table 1. Range of presentations

Presenting problem	Number of recordings
Depression	9
Panic disorder	4
Phobia	3
Generalized anxiety disorder	3
Obsessive compulsive disorder	3
Social anxiety	2
Health anxiety	1

Table 2. Therapy sessions from which recordings were rated

Session number	Number of occurrences
3	2
4	6
5	1
6	7
7	4
8	3
9	1
10	1

Session recordings

Twenty-five video recordings (each approximately 50 minutes in duration) were identified for rating by an independent research assistant. These were selected purposively rather than randomly from the 76 video recordings submitted as part of the standard schedule of assessments on the training programme (a decision that may have introduced bias, but which allowed for the characteristics of recordings to be monitored and so balanced). All were early or mid-treatment sessions (with assessments and final sessions excluded on the grounds that these (by definition) will have a restricted focus). As the study progressed the range of presenting problems was balanced, so as to ensure that there was reasonable representation of different disorders (see Table 1). Most recordings were taken from the middle stage of therapy, with only a minority from the initial or final stages of the intervention (Table 2). Eight therapists contributed a single recording, one contributed two recordings and five contributed three recordings.

Raters

Six raters contributed to the study; all were employed as tutors on an IAPT Postgraduate Diploma programme, and so routinely appraised the work of trainees as part of the examination process. All were female Clinical Psychologists accredited as CBT therapists with the British Association for Behavioural and Cognitive Psychotherapies (BABCP), and had considerable experience both as clinicians (range 6–14 years, mean 7.6 years) and as tutors with the programme (range 3–6 years; mean 4 years). Raters reviewed all 25 recordings independently, and so were blind as to the ratings of their colleagues.

Rating scales

Each rater evaluated the whole sample of recordings using (a) the Cognitive Therapy Scale-Revised (Blackburn *et al.*, 2001) and (b) the UCL generic and CBT competence scales (as described in Roth, 2016).

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

Table 3. Intraclass correlation coefficients on the UCL generic and CBT scales and the Cognitive Therapy Rating Scale

	ICC for all raters (95% confidence intervals)	ICC with outlier removed (95% confidence intervals)
UCL CBT scale	0.394 (0.228–0.598	0.476 (0.294–0.657)
UCL generic scale	0.272 (0.126–0.478)	0.346 (0.174–0.562)
CTS-R	0.424 (0.260–0.621)	0.516 (0.339–0.702)

Training of raters

CTS-R

All six raters had received extensive training in the use of the CTS-R as part of their work with the programme, including annual consensus and review meetings aimed at ensuring consistency in their scoring. As such, they were not offered further training on this instrument.

UCL competence scales

There was limited training in the use of the UCL scales. As noted earlier, the intent was to approximate 'real-world' application of the scale, and so rely on the instruction materials accompanying the scale and the scale itself. Training consisted of a meeting with all six raters focused on an initial session rating, allowing the opportunity for feedback on the scale itself and identifying any areas which were ambiguous or required clarification. A further mid-point consensus meeting was held after 10 session ratings had been completed; a previously rated recording (subsequently excluded from the study) was reviewed, and clarification of the rating system discussed. This was followed by a 'live' rating of a further session (again, excluded from the study), which allowed for group discussion of reasons for any variation in scoring.

Controlling for order effects

The order in which the CTS-R or the UCL scales were applied was balanced both across recordings and across raters, so as to mitigate the risk that rating on one scale could influence ratings on the other.

Exploring variations in ratings

Raters were asked to indicate the reasons for each rating; additionally (at the mid-point of the study), raters and investigators met to discuss progress and to undertake a joint rating session in which all raters viewed a session recording and rated it independently before sharing their ratings and discussing reasons for variance. Taken together, these sources provided information about the decision-making processes associated with the scale.

Results

In this analysis the intra-class correlation coefficient (ICC) is computed for absolute agreement and for single raters using a two-way mixed effects model; results are displayed in Table 3.

UCL CBT scale

Across the six raters, the ICC for the scale total was poor to moderate (ICC = 0.394: 95% confidence interval 0.228–0.598).The mean correlation between raters was 0.45 (range 0.26–0.74). As can be seen from Table 4, one rater had consistently low correlations with the other raters:

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

Table 4. Correlations between raters on each scale

CTSR					
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 2	0.52				
Rater 3	0.29	0.13			
Rater 4	0.50	0.50	0.36		
Rater 5	0.52	0.61	0.12	0.59	
Rater 6	0.49	0.45	0.14	0.66	0.44
UCL generic scale					
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 2	0.42				
Rater 3	0.26	−0.10			
Rater 4	0.30	0.29	0.38		
Rater 5	0.40	0.57	−0.11	0.21	
Rater 6	0.49	0.40	0.31	0.59	0.15
UCL CBT scale					
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 2	0.46				
Rater 3	0.35	0.09			
Rater 4	0.63	0.51	0.46		
Rater 5	0.61	0.58	0.25	0.54	
Rater 6	0.48	0.53	0.36	0.74	0.25

removing this individual from the analysis increased the ICC to 0.476 (95% confidence interval 0.294–0.675), and the mean correlation between raters to 0.52.[4]

UCL generic scale

The ICC for the total scale was poor (ICC = 0.272: 95% confidence interval 0.126–0.478), with a mean correlation between raters of 0.32 (range −0.11 to 0.57). The same rater had consistently low correlations with the other raters: removing them from the analysis increased the ICC to 0.346 (95% confidence interval 0.174–0.562), and the mean correlation between raters to 0.43.

CTS-R

The ICC for the total scale was poor to moderate (ICC = 0.424: 95% confidence interval 0.260–0.621), with a mean correlation between raters of 0.44 (range 0.12–0.67). Once again the same rater had consistently low correlations with their colleagues: removing this individual from the analysis increased the ICC to 0.516 (95% confidence interval 0.339–0.702), and the mean correlation between raters to 0.56.

Exploring variations in ratings

All raters were asked to explain each of their rating decisions, and also undertook a joint rating session in which they viewed a session recording and rated it independently before sharing their ratings and discussing reasons for variance. This made it possible to explore some of the reasons for discrepant ratings and the dilemmas that raters attempted to resolve. A systematic analysis is out of the scope of this paper, but some examples are helpful:

[4] Removing the outlier rater represents a *post-hoc* exploratory analysis. It is relevant that variance between this and other raters was also apparent when session recordings submitted as part of the training programme were rated (and was subsequently addressed). Identifying an empirical reason for this discrepancy is not possible, but its consistency across different contexts does provide some justification for the exploratory analysis.

(a) Faced with significant intra-session variation in competence (with a specific skill being applied well or poorly at different points) raters sometimes awarded an averaged score or rated in line with the best or poorest examples.

(b) Applying CBT techniques requires attention both to structure (how something is set up) as well as content (identifying and working with material that is salient). Therapists sometimes employed a technique (such as evaluating negative automatic thoughts, or setting up a behavioural experiment) in a way that was appropriately structured but focused on content that was not central to the client's issues (so reducing its potential value). Some raters responded by awarding a low rating (noting that the content was misjudged), whereas others awarded a high rating on the basis that the therapist had set up the technique in a skilful manner.

(c) On occasion raters identified significant clinical themes that had not been picked up by their colleagues (for example, noting that the therapist had missed an important issue), and this led to their appraising specific techniques differently from other raters.

Discussion

The present study set out to contrast the UCL scales against the CTS-R, hypothesizing that levels of inter-rater reliability on all three scales will be low. In summary, the inter-rater reliability of the UCL generic and CBT scales was poor (with ICCs of 0.272 and 0.394, respectively). Removing the rater with a consistently low level of agreement with their colleagues improved the ICCs for the CBT scale closer to a moderate level (0.476). While the ICC for the generic scale also improved, it remained poor at 0.346. Scores for the CTS-R were broadly comparable to that on the UCL CBT scale; based on ratings from all raters the ICC was poor (ICC = 0.424) , with a moderate level of reliability if the outlier rater is removed (ICC = 0.516).

These estimates for the reliability of the UCL scales are low, although they are comparable to the lower end of the range of ICCs found in studies of the CTS-R (Loades and Armstrong, 2016). The raters all had considerable experience using the CTS-R to evaluate session recordings of trainees on the same training programme, with periodic meetings aimed at checking the consistency of their ratings. In contrast, training on the UCL scales was minimal and restricted to an initial meeting at which the scales were discussed, rating of an initial recording, and a concordance meeting after 10 recordings had been evaluated. The figures for the CTS-R therefore represent a benchmark, against which the UCL CBT scale performs equivalently.

These findings confirm that there are significant difficulties in achieving high levels of reliability in whole-session structured assessment of therapist competence. Disparities in ratings could be attributable to any number of causes, among the most basic being deficiencies in the way the scale is structured or ambiguity in scale descriptors. Rating therapist competence is inherently challenging: while manuals can attempt to anchor ratings, unless scale items are very specific and straightforward, raters will inevitably apply idiosyncratic clinical judgements regarding the 'meaning' of a scale item, and so arrive at different (but legitimate) ratings. Some of the reasons for discrepant ratings and the dilemmas that raters attempted to resolve have been noted above; each example illustrates a challenge to interpretation of scale items, despite the fact that each item was anchored with several examples of the behaviours and actions associated with each area of competence. However, not every eventuality can be anticipated, and at points raters will inevitably fall back on idiosyncratic conceptions. It may also be significant that the outlier rater was a temporary member of staff with the training programme, and not subject to the marker training and inter-rater checks that applied to the other raters. This may speak to the value of such procedures in constraining rater 'drift'.

One solution to the concerns raised above is to recognize that rating specific competences is an inherently complex and potentially unstable task because of the number of variables that need to

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

be accounted for. Recognizing this, some authors (e.g. Elkin, 1999) have suggested that there may be advantages to global rather than specific rating scales. Kuyken and Tsivrikos (2004) developed a four-item scale that rated therapists' overall competence, overall skills in CBT, their flexibility, and their general skills, finding that all the scale items were highly inter-correlated, and significantly associated with outcome. The risk with this approach is that it is impossible to know how each rater arrives at a judgement; as such even if their overall ratings are congruent they might be based on different criteria. A workable compromise might be to combine the two approaches, with a scale that asks for global judgements based on detailed descriptions of specific therapist behaviours (as exemplified, for example, by the Assessment of Core CBT Skills (Muse *et al.*, 2017).

A significant limitation to this study is the fact that the performance of the scale was only examined in relation to therapists in training. This is likely to have constrained the range of scores (as fewer items would be rated as fully competent). There are also no data on test–retest reliability, or the degree to which the scale is sensitive to training effects. Given the ways in which the scale is likely to be deployed, this latter limitation is something that future studies may wish to examine.

Conclusions

Results from this study raise doubt about the capacity of raters to score structured competence items reliably in contexts where there is minimal opportunity for them to calibrate their scores through a training programme that aims to achieve consensus. This level of uncertainty matters less when using the scales for formative assessments, as the feedback on areas requiring improvement would still be useful. However, their use as a summative evaluation of competence is not supported without additional measures that can help to triangulate the assessment (such as extensive reliability checks, blind double-marking, moderation and external examiners and reports of direct observation from supervisors).

Acknowledgement. The authors would like to thank Glenys Parry and Roz Shafran for their helpful comments on an earlier version of this paper, and the raters involved in this study: Anna Coughtrey, Jenny Griggs, Maria Loades, Natalie Peace, Melissa Snaith and Hannah Whitney.

Financial support. This study was funded by NHS Education for Scotland.

Conflicts of interest. Anthony Roth, Pamela Myles-Hooton and Amanda Branson have no conflicts of interest with respect to this publication.

Ethical statement. The authors have abided by the Ethical Principles of Psychologists and Code of Conduct as ethical approval for this study was granted by the UCL Research Ethics Committee (CEHP/2013/507).

References

Beck, A. T., Rush, A. J., Shaw, B. F. and Emery, G. (1979). *Cognitive Therapy of Depression*. New York, USA: Guilford Press.

Blackburn, I. M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A. and Reichelt, F. K. (2001). The Revised Cognitive Therapy Scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy*, *29*, 431–446. doi: 10.1017/S1352465801004040

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., et al. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology*, *74*, 658–670. doi: 10.1037/0022-006X.74.4.658

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: disentangling therapists from therapies. *Clinical Psychology: Science and Practice*, *6*, 10–32. doi: 10.1093/clipsy.6.1.10

Jacobson, N. S. and Gortner, E. T. (2000). Can depression be de-medicalized in the 21st century: scientific revolutions, counter-revolutions and the magnetic field of normal science. *Behaviour Research and Therapy*, *38*, 103–117. doi: 10.1016/S0005-7967(99)00029-7

Kuyken, W. and Tsivrikos, D. (2004). Therapist competence, comorbidity and cognitive-behavioral therapy for depression. *Psychotherapy and Psychosomatics*, *78*, 42–48. doi: 10.1159/000172619

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press

Loades, M. E. and Armstrong, P. (2016). The challenge of training supervisors to use direct assessments of clinical competence in CBT consistently: a systematic review and exploratory training study. *The Cognitive Behaviour Therapist*, *9*, e27. doi: 10.1017/S1754470X15000288

McManus, F., Rakovshik, S., Kennerley, H., Fennell, M. and Westbrook, D. (2012). An investigation of the accuracy of therapists' self-assessment of cognitive-behaviour therapy skills. *British Journal of Clinical Psychology*, *51*, 292–306. doi: 10.1111/j.2044-8260.2011.02028.x

Muse, K. and McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review*, *33*, 484–499. doi: 10.1016/j.cpr.2013.01.010

Muse, K., McManus, F., Rakovshik, S. and Thwaites, R. (2017). Development and psychometric evaluation of the Assessment of Core CBT Skills (ACCS): an observation-based tool for assessing cognitive behavioural therapy competence. *Psychological Assessment*, *29*, 542–555. doi: 10.1037/pas0000372

Roth, A. D. (2016). A new scale for the assessment of competences in cognitive and behavioural therapy. *Behavioural and Cognitive Psychotherapy*, *44*, 620–624. doi: 10.1017/S1352465816000011

Roth, A. D. and Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *36*, 129–147. doi: 10.1017/S1352465808004141

Waltz, J., Addis, N. E., Koerner, K. and Jacobson, N. (1991) Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, *61*, 620–530. doi: 10.1037/0022-006X.61.4.620

Young, J. and Beck, A. T. (1980). Cognitive Therapy Scale: Rating Manual (unpublished manuscript).

Cite this article: Roth AD, Myles-Hooton P, and Branson A (2019). Judging clinical competence using structured observation tools: A cautionary tale. *Behavioural and Cognitive Psychotherapy* **47**, 736–744. https://doi.org/10.1017/S1352465819000316

https://doi.org/10.1017/S1352465819000316 Published online by Cambridge University Press