Development of a General Legal Confidence Scale: A First Implementation of the Rasch Measurement Model in Empirical Legal Studies

*P. Pleasence and N.J. Balmer**

ABSTRACT

Legal capability has long been of evident importance in our understanding of legal problem resolution behaviour. Although legal capability remains a contested concept, there is much commonality between specifications. Some aspects are generic, while others – such as legal confidence – are particular to law. Such law specific measures as have been developed to date have been developed in an ad-hoc fashion; with no attempts made to test psychometric properties, using either classical test theory or modern psychometric methods. This has been a shortcoming in the empirical legal field, weakening theoretical development and precluding reliable estimation of changes in levels of legal capability over time. In this paper, we set out details of a study aimed at introducing new methods to scale development in the field of empirical legal studies; based on the approaches that have evolved in other fields and the latest developments in psychometric modelling. Specifically, we set out details of the use of a specially designed item pool – based on an increasingly demanding legal scenario – and Rasch analysis to develop a general legal confidence scale. Once the twelve item pool items were reduced to a final set of six, this yielded a scale with good psychometric properties: a General Legal Confidence (GLC) Scale. The scale showed good overall fit, item fit, person fit, targeting and internal consistency. All items had ordered thresholds, there was no response dependence, items were unidimensional and there was no evidence of differential item functioning. The GLC scale constitutes an effective measure of general legal confidence, and demonstrates it is possible to arrive at robust and coherent law specific measures of legal capability through the careful design of questions and application of the latest psychometric modelling techniques.

I. INTRODUCTION

Galanter (1976:936) argued that "lack of capability poses the most fundamental … barrier to access [to legality]," and Felstiner et al. (1981) that the ability to 'name', 'blame' and 'claim' is central to the emergence and transformation of legal disputes. Thus, legal capability has long been of evident importance in our understanding of legal problem resolution behaviour. Moreover, continuing processes of 'juridification',[1] the emergence of disruptive technologies and new forms of legal service delivery,[2] limits to legal aid,[3] policy and operational refocusing on legal service users' (rather than providers') needs,[4] and generally greater understanding of

* Address correspondence to Professor Pascoe Pleasence, Faculty of Laws, University College London, Bentham House, Endsleigh Gardens, London, WC1H 0EG, UK; email: p.pleasence@ucl.ac.uk. Professor Balmer is at the Faculty of Laws, University College London.

We are very grateful to The Legal Education Foundation, and in particular Matthew Smerdon and Dr Natalie Byrom, for funding the research and for their continued support for our work. We would also like to thank the editor, Dawn Chutkow, and anonymous referees for their carefully considered and thorough reviews, which significantly enhanced the research.

[1] Defined by Habermas (1987:357) as "the tendency towards an increase in formal (or positive, written) law that can be observed in a modern society."

[2] See, for example, Susskind (2008), Smith (2014).

[3] Or even cuts, such as those brought about by the Legal Aid, Sentencing and Punishment of Offenders Act 2012 in England and Wales.

[4] See, for example, Legal Services Commission (2006), Commonwealth Attorney-General's Access to Justice Taskforce (2009), Canadian National Action Committee on Access to Justice in Civil and Family Matters (2013).

factors influencing legal problem resolving behaviour,[5] have acted to increase interest in both understanding legal capability and, beyond that, measuring it.

While the concept of legal capability remains contested, it links closely to Sen's (1999, 2002, 2010) capability approach to disadvantage and refers to the capabilities necessary for a person "to resolve legal problems effectively" (Coumarelos et al. 2012:29). Legal capability is therefore multi-dimensional; incorporating a range of more narrowly framed capabilities, across a variety of domains.

Although specifications of legal capability have differed, there is much commonality. For example, Parle's (2009), Collard et al.'s (2011) and Pleasence et al.'s (2014) specifications all incorporated knowledge of law, the ability to spot legal issues, awareness of legal services, understanding of and the ability to assess dispute resolution options, planning and management skills, communication skills, confidence and emotional fortitude; within a series of similarly formulated domains.[6] There is also much commonality between specifications of legal capability and those of (more narrowly construed, but conceptually similar) lawyer competence.[7] Evidently, some aspects of legal capability – such as communication skills – are largely generic, while others – such as confidence in one's ability to effectively address 'justiciable'[8] problems – are particular to law. In relation to many generic aspects of legal capability, robust measures have been developed in other fields. However, an enormous challenge remains to develop appropriate law specific capability measures. Such measures of

[5] See, for example, Pleasence and Balmer (2014), Pleasence, Balmer & Denvir (2015).

[6] Parle's (2009) small-scale qualitative study of young people's legal capability suggested six core domains: knowing rights and remedies; spotting a legal issue; knowing where to go for help; planning how to resolve the issue; communicating effectively; and managing emotions. Collard et al.'s (2011) later review listed 22 separate aspects of legal capability across four similarly defined domains: recognising and framing the legal dimensions of issues and situations (including knowledge of law and the ability to frame and explain situations in terms of the law); finding out more about the legal dimensions of issues and situations (including the ability to identify assess sources of legal information and dispute resolution processes); dealing with law-related issues (including confidence and the ability to communicate, plan, manage and assess processes and outcomes); and engaging and influencing (including awareness of means to achieve change and the confidence and strength to effect change). More recently still, Pleasence et al. (2014) detailed a similar range of aspects, albeit within more abstract and broadly recognised domains: knowledge (of, for example, law, legal services and dispute resolution options); skills (including communication skills and the ability to identify legal issues); attitudes (to, for example, law itself, legal services, legal institutions and legal systems); attributes (such as confidence, emotional fortitude and preparedness to act); and resources (including financial, technological and social). Similar, though less extensive, conceptualisations have also been offered by, for example, Balmer et al. (2010), Coumarelos et al. (2012) and the Canadian National Action Committee on Access to Justice in Civil and Family Matters (2013). For example, the Canadian National Action Committee on Access to Justice in Civil and Family Matters (2013) defined legal capability as "the level of knowledge, skills and confidence as well as the attitudes of people that allow them to: recognize that there are legal components or aspects to many activities and events of everyday life; better anticipate and manage these components; be able to sort legal from non-legal aspects of their problems and address their interdependence; avoid unnecessary escalation of conflicts into more serious problems or disputes that may require legal intervention; assess options that are available and that foster reasonable solutions in situations of conflict; be aware of when and how legal representation can assist with disputes and how to access legal representation." See, also similar conceptualisations in relation to financial capability (e.g. Kempson et al. 2005).

[7] For example, Sherr at al (1994) summarised the elements of competent lawyering as being: legal knowledge; practical skills (e.g. interviewing, negotiating, drafting); administrative skills (e.g. practice management, supervision); motivation (to perform in an effective manner); proficiency to plan and prepare; mental and physical faculties; and understanding of limitations. Similarly, Shultz and Zedeck's (2011) extensive study identified aspects of lawyer effectiveness to include: intellectual and cognitive (e.g. analysis, reasoning, problem solving, judgement); research and information gathering (e.g. researching the law, fact-finding, and questioning); communication (e.g. influencing, writing, speaking, listening); planning and organisational (e.g. strategic planning, self-management); conflict resolution (negotiation, empathy); character (e.g. diligence, stress management).

[8] Justiciable problems are problems that raise legal issues, whether or not this is recognised by those facing them, and whether or not any action taken to deal with them involves lawyers or the wider legal system (Genn 1999).

legal capability as exist have been developed in an ad-hoc fashion; with no attempts made to test psychometric properties, using either classical test theory or modern psychometric methods. This is a major shortcoming. At a time when the profile of access to justice policy has never been higher – with the adoption of United Nations Sustainable Development Goal 16.3 (encompassing the target of access of justice for all)[9] and growing recognition of links between access to justice and inclusive and sustainable growth (e.g. OECD & Open Society Foundations 2016) – the absence of robust measures of legal capability hampers the investigation of increasingly evident links between legal capability and legal problem resolution behaviour (e.g. Sandefur 2007; Gramatikov & Porter 2011, Pleasence & Balmer 2014, Pleasence et al. 2015), and precludes effective quantitative outcome evaluation of efforts to increase legal capability among, particularly disadvantaged, populations (such as, in the United Kingdom, the *Making Our Rights Reality* initiative led by Youth Access, or the work of *Law for Life* and *Young Citizens* more generally[10]). The absence of robust measures also weakens theoretical development in the area of legal capability. But, unfortunately, there is no tradition of standardised measurement within the field, as there is in larger and often more technically robust fields of social science research such as health and psychology.[11]

A. Measuring Legal Confidence

A number of empirical legal studies have incorporated measures of legal confidence; notably, at least 15 of the more than 50 large-scale stand-alone national legal needs surveys conducted around the world over the past 25 years.[12] In most cases, the focus has been on general legal confidence; although in two cases it was instead on confidence in relation to the resolution of specific problems reported by respondents.[13]

The majority of surveys have incorporated variants of the 'subjective legal empowerment' questions developed by Gramatikov and Porter (2011) and routinely used within the Hague Institute for the Internationalisation of Law's (HiiL) *Justice Needs and Satisfaction Surveys*.[14] Gramatikov and Porter (2011:169) described subjective legal empowerment as "the subjective self-belief that a person possesses … [in their] ability to mobilise the necessary resources, competencies, and energies to solve particular problems of a legal nature."[15] The concept was conceived as a domain specific form of self-efficacy; defined by Bandura (1997:3) as referring to "beliefs in one's capabilities to organise and execute the courses of action required to produce given attainments."

Gramatikov and Porter (2011) style questions involve asking respondents about the likelihood of their being able to achieve a fair solution (or, separately, a solution and a solution that is fair) to a dispute; with questions typically asking about problems involving six distinct

[9] United Nations General Assembly Resolution 70/1, 25th September 2015. The United Nations Sustainable Development Goals build on the earlier Millennium Development Goals and are collectively directed towards ending poverty, ensuring economic prosperity and sustaining the environment.

[10] See, for example, http://www.lawforlife.org.uk/public-legal-education/ and https://smartlaw.org.uk/lawyers-in-schools/ (accessed on 12th March 2018).

[11] Some of the standardised measures developed within the fields of health and psychology have been used in legal need surveys (e.g. SF-12 (Ware, Kosinski & Keller 1996), GHQ-12 (Goldberg & Williams 1988), elements of the Big Five Inventory (John, Naumann & Soto 2008)), and significant development work preceded the introduction of a standardised measure of problem severity to the 2010 *English and Welsh Civil and Social Justice Panel Survey*. However, beyond this, standardised measurement has been largely limited to the use of 'harmonised' (e.g. Office for National Statistics 2015) demographic questions.

[12] Spanning in excess of 30 jurisdictions. Pleasence & Balmer (forthcoming).

[13] the 2005 Japanese (Murayama 2007) and 2012 Tajik surveys (Social Research Center 2012).

[14] See, for example, HiiL (2014)

[15] Within the socio-legal tradition, the concept of subjective legal empowerment is rooted in the broader concept of legal empowerment, expounded by Golub and McQuay (2001) in the law and development context.

types of issue. The inclusion of a range of issue types is aimed at addressing possible differences in levels of self-efficacy between them. As Sander and Sanders (2003:3) succinctly put it, in the context of academic study, "people have different levels of confidence in different situations."[16]

Responses to Gramatikov and Porter (2011) style questions appear to link to problem resolving behaviour. Analysis of a set of questions included in the 2012 *English and Welsh Civil and Social Justice Panel Survey* indicated that "as respondents' subjective legal empowerment scores increased, inaction significantly decreased" (Pleasence & Balmer 2014:32). This is not surprising. As Gramatikov and Porter (2011:172) observed, "power is the currency of disputes;" with dispute resolution often occurring, as Galanter (1974) described, within a framework of myriad power imbalances. Moreover, the broader self-efficacy literature points to confidence being an important influence on behaviour across a range of domains, including health behaviour (e.g. Strecher et al. 1986, Grembowski et al. 1993), career development (e.g. Dawes et al. 2000) and athletic performance; where "research has demonstrated self-confidence to be one of the most influential cognitive determinants of athletic performance" (Beattie et al. 2011:184).[17]

However, while Gramatikov and Porter's approach has yielded results of some interest, is carefully considered and has a solid theoretical grounding, they themselves noted that measuring self-efficacy in relation to law is "a challenging endeavour,"[18] and neither classical test theory nor modern psychometric methods have been used in question (or scale) development. In fact, subjective legal empowerment questions have tended to exhibit a number of structural weaknesses; particularly concerning their broad formulation.

The questions used in the 2012 *English and Welsh Civil and Social Justice Panel Survey* were particularly broad – asking about the likelihood that respondents "would be able to get a fair solution" if they had a conflict "with your employer," "with a family member," "with a neighbour" and about a "land dispute," "business dispute" and if respondents "became a victim of crime," with no further details provided – and later cognitive testing of similar questions highlighted that they were 'not clear', 'not specific enough' and 'too complicated'.[19] And although formulations used within the *Justice Needs and Satisfaction Survey* have included greater detail,[20] the subject matters of the questions are evidently open to a broad range of interpretations.

Moreover, as was recognised by Gramatikov and Porter (2011), but not operationalised within their questions,[21] self-efficacy is multidimensional within the context of different issues. As was made evident in the specifications of legal capability referenced above, in resolving a particular type of legal problem, an individual may need to, among many other things, accrue information, communicate and negotiate with another party, navigate a formal process, and present a case, each of which may attract very different levels of confidence. Furthermore, as Bandura (1997:36) expounded, "efficacy beliefs are concerned not only with the exercise of control over action but also with the self-regulation of thought processes, motivation, and

[16] Thus, self-efficacy is likely to vary between such activities as, for example, distance learning (Tang & Tseng 2013) and mathematics (OECD 2003).

[17] Including, famously, in arm wrestling (Nelson & Furst 1972).

[18] And it is always important to be mindful of Cameron's (1963, p.13) caution that "not everything that counts can be counted."

[19] As part of the development of capability measures of young people attending Youth Information, Advice and Counselling (YIAC) services.

[20] HiiL (2014)

[21] Space within questionnaires for particular topics is often very limited. Gramatikov and Porter (2011) were legitimately concerned to incorporate only a few short questions.

affective and physiological states;"[22] each reflecting further self-efficacy domains.

To an extent, this multidimensionality of legal confidence was reflected in Mackie's (2013:10-11) operationalisation of Collard et al.'s (2011) capability framework for the *Legal Capability for Everyday Life* evaluation. This included questions about confidence in relation to understanding legal rights and obligations and, separately, in relation to knowing when to get expert help to deal with a situation.

Of course, the multidimensionality of legal confidence does not necessarily preclude broad measurement. Nor does it mean that confidence in addressing justiciable problems simply amounts to the sum of levels of confidence in relation to each of these factors. As Bandura (1997:50) explained, "a multidimensional approach does not mean that there is no structure or generality to efficacy beliefs. The development and exercise of capabilities would be severely constricted if there was absolutely no transfer of efficacy beliefs across activities or settings."[23]

In this paper, we set out first details of a study aimed at introducing new methods to scale development in the field of empirical legal studies; based on the approaches that have evolved in other fields (e.g. DeVellis 2012) and the latest developments in psychometric modelling (Hobart & Cano 2009, Christensen et al. 2013, Boone et al. 2014, Bond & Fox 2015). Specifically, we set out details of our development of a general legal confidence scale.

B. Approaches to Scale Development

There are a broad range of disciplines, including education, health and psychology, where there is interest in, and extensive experience of, measuring personal attributes that cannot be directly observed; such as intelligence, self-esteem or anxiety. Across these disciplines, principles and methods of robust scale development have evolved.

Scale development typically involves a number of set steps, as set out in DeVellis (2012). These include determination of what is to be to measured, generation of an item pool (in our case, a series of statement-based questions with a Likert scale response format), determination of the measurement format, expert item pool review, consideration of inclusion of validation items, administration of items to a development sample, evaluation of items and optimisation of scale length. The final two steps in this process typically involve either classical test theory (CTT) or modern psychometric methods (in our case Rasch analysis).

1. Classical test theory (CTT) and modern psychometrics

Until recently, the most widely used method for constructing and evaluating rating scales was what is commonly termed traditional psychometric methods, underpinned by CTT methods (sometimes also referred to as true score theory) (e.g. DeVellis 2006, 2012). In CTT, each person is assumed to have an observed score (O) that represents their true score for a trait (T) plus an error term (ε), with $O = T + \varepsilon$. The standard deviation of the errors (the standard error of measurement) is directly related to reliability, with reducing errors moving the observed

[22] Thus, for example, Vealey and Chase (2008) identified three kinds of confidence pertaining to sport (cognitive efficiency, physical skills and training, resilience) and Schwarzer and Fuchs (1995) identified five kinds of confidence pertaining to addictive behaviours, such as smoking (resistance, harm reduction, action, coping, recovery).

[23] Bandura (1997:51) argued that this can occur through five means: where there are similar sub-skills; co-development of skills (as in a school environment where many subjects are learned in parallel and general perceptions of efficacy may develop); where proficiency depends upon "selecting and orchestrating subskills guided by higher self-regulatory skills;" generalisability of coping skills; and where commonalities between activities are highlighted in learning.

scores closer to the true scores (Smith et al. 2016). Importantly, CTTs psychometric properties are at an overall test level rather than at an individual item level. Error scores are assumed to be uncorrelated with each other and with the true scores, and observed and true scores are linearly related. However, since true scores and error scores cannot be determined, the appropriateness of the assumptions cannot be verified (Allen & Yen 2002). In addition, many rating scales (including the majority of our items) employ Likert scale type response formats (with sequentially ordered response options assigned sequentially ordered integers). Traditional methods assume that ordinal level total scores approximate to interval level measures (Allen & Yen 2002), which is not the case and should not be treated as such (Smith et al. 2016). Other problematic assumptions include a homogeneous contribution of items to the final score and equivalence of response options among different items (Martinez-Martin & Forjaz 2012). Moreover, with traditional methods, evaluations of scales are sample dependent and the measurement of people is scale dependent, undermining the use of such total scores as measurements (Allen & Yen 2002).

Subsequently, modern psychometrics were developed (Item Response Theory (IRT) and Rasch analysis). Like CTT, the methods set out theories of how rating scale scores relate to measurements of the variables they seek to estimate (e.g. legal confidence). However, unlike CTT, they are underpinned by mathematical models of the theories, enabling their verification through formal and rigorous testing (Allen & Yen 2002). Modern psychometrics (which include IRT and Rasch models), have become standard in many fields, with Rasch analysis perhaps the most widely used method for scale development/item reduction (with some suggestion that it is the only method specifically developed for constructing measurement (Smith et al. 2016)).[24] The Rasch model is set out below, though the basic rationale is that an individual's response to a specific item is based on (a log function of) their characteristics (e.g. their ability - such as legal confidence) and characteristics of the item (i.e. its difficulty, or level of confidence required to endorse it). So, for Rasch analysis, psychometric properties are at the item, rather than the test level.

While the Rasch model is identical mathematically to a one parameter logistic model in IRT, they do differ. In IRT, if items do not fit the model, the aim would be to find a more suitable model. In Rasch analysis, the aim is to examine fit and anomalies and then adapt the data to the Rasch model to create a more valid and reliable instrument (Smith et al. 2016). Andrich (2004) provides a detailed contrast of IRT and Rasch models. Unlike CTT and IRT, Rasch analysis can produce sample free and test free measurement (Smith et al. 2016, Hays et al. 2000), with item difficulty the same regardless of who is in the sample and person ability estimates the same regardless of which items are included. This unique and useful property of Rasch is called specific objectivity, allowing invariant measurement (Engelhard 2013, Smith et al. 2016, Hays et al. 2000) with person and item parameters separable and measured on the same invariant log scale.[25]

More generally, unlike CTT, modern methods (including Rasch) can also explore whether items are equivalent in meaning to different respondents (differential item functioning – described in detail below), allow inclusion of items with different response formats on the

[24] E.g. Blanchin et al. (2010) noted that interesting psychometrics properties including the exhaustivity of the score on the latent trait and the specific objectivity has driven much of the increased use of Rasch in recent years, with Hays and Lipscomb (2007) and Smith et al. (2016) also noting growth in use in health.

[25] The significance of the property of invariance has been challenged by some authors, with Fan (1998: 361) suggesting, "the superiority of IRT over CTT in this regard has been taken for granted by the measurement community, and no empirical scrutiny has been deemed necessary". Streiner et al., (2015) also noted studies indicating large differences in measurement from one population or test condition to another, suggesting cases where invariance may not hold. There has been some suggestion that disagreement regarding the invariance of item characteristics may be related to the population studied (Streiner et al., 2015) and how homogeneous/heterogeneous they are (Cella and Chang, 2000).

same scale, assess person fit (e.g. assessment of the extent to which individual respondents provide useful data or are taking the exercise seriously – further detail below) and allow computer adaptive testing (Hays et al. 2000). Missing data is also less of an issue for Rasch analysis (Boone et al. 2014). As psychometric properties are at an item level (unlike CTT), individual items can be comprehensively evaluated. Crucially, scales developed using Rasch analysis also allow ordered observations (such as Likert scales) to be transformed into an interval scaled measure of the latent trait (Salzberger 2010, Wright & Linacre 1989), allowing for a broader range of statistical analyses (see Wright & Linacre 1989).[26]

An introduction to Rasch analysis and its use can be found in Boone et al., (2014) or Bond and Fox (2015), while Tennant and Conaghan (2007) provide a helpful guide to using, and reporting findings from Rasch analysis. Hobart and Cano (2009). Smith et al. (2016), Wright (1992) and Hays et al. (2000) also provide a more detailed contrast of Rasch analysis, other modern forms of psychometric analysis and traditional methods.[27]

2. The Rasch Model

The basic Rasch assumptions are that: (a) each person is characterized by an ability, and (b) each item by a difficulty that (c) can be expressed by numbers along one line. Finally, (d) from the difference between the numbers (and nothing else), the probability of observing any particular scored response can be computed (Bond & Fox 2015). Equation 1 shows the basic Rasch model.

Equation 1. The Rasch model

$$p_{ni} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

Where p_{ni} = the probability of affirming (i.e. giving a positive response), for item i and person n, D_i = the difficulty of item i, B_n = the ability of person n

Rasch is a logistic model (i.e. the expression e/(1+e) is central). Put simply, in our context, the probability of affirming an item is a logistic function of the difference between an individual's confidence and the level of confidence an item expresses. So, where an individual's confidence is high and level of confidence expressed by a positive response to an item is low (i.e. an easy item), probability tends towards one. Conversely, where an individual's confidence is low and level expressed by a positive response to an item is high, probability tends towards zero.

II. METHODS

[26] i.e. they are not nominal or ordinal. For example, conducting parametric analyses with ordinal data assumes equal intervals between successive ranks (e.g. on a Likert scale, see Martinez-Martin & Forjaz (2012)). Rasch analysis addresses this problem.

[27] Despite the well documented advantages of Rasch analysis, it is not without its critics, notably Goldstein (1979, 2010, 2015). This includes the suggestion that fitting the data to a model (Rasch) rather than a model to the data is a radical proposal. For further discussion, see Bond and Fox (2015) and Linacre and Fisher (2012). Moreover, despite the documented advantages of Rasch (and IRT methods), differences between scales constructed with IRT/Rasch and CTT are often trivial, especially with large sample sizes (Fan, 1998; Streiner et al., 2015), with studies employing both modern and classical methods often reporting high degrees of association between scale scores (e.g. Prieto et al., 2003; Petrillo et al., 2015).

Two dedicated general population probability surveys were conducted to facilitate the development of a general legal confidence scale. In addition to general legal confidence questions, they also included a range of additional questions on narrower aspects of legal confidence, attitudes to justice and experience of law. The first survey acted as a pilot, to test initial sets of legal confidence questions; the most promising of which were then supplemented in the second survey. One of these sets of questions – detailed below – was expanded into a 12-item pool within the second survey, specifically directed towards measuring general legal confidence.

A. Survey Administration

Both surveys utilised a one-stage sample design in which a stratified, but unclustered, sample of addresses was drawn from the Residential Postcode Address File; the cornerstone of national probability samples in England and Wales. The survey was an innovative hybrid form of postal and online survey, based on the Community Life Survey web experiment (TNS-BMRB 2013).[28] The findings of the Community Life Survey web experiment suggest that the quality of data obtained from hybrid online and postal surveys can generally be expected to be similar to that which would have been obtained from face-to-face surveys.[29] Advance letters (and reminder letters and postcards) invited "the person aged 16 or over who has the next birthday" in a household to either take part in the survey online (using a provided web-link) or to return a postal version of the questionnaire, if requested. The person completing the survey was offered a £10 shopping voucher as an incentive.

There were 1,061 respondents to the second survey, of whom 872 completed all the sections.[30] The second survey yielded a broad range of socio-demographics, which was the key consideration for the scale development exercise.[31]

B. *Progressing Scenario General Legal Confidence Questions*

[28] The Community Life Survey web experiment was "one of the largest ever tests of web survey methodology in which random sampling has been employed" (TNS-BMRB 2013, p.4).

[29] As the authors of the Community Life Survey web experiment noted, although there were notable differences observed between the face-to-face and online-postal samples, web/postal respondents took the same length of time to complete the survey questionnaire and generally yielded similar estimates; although estimates from the different variants varied and it was suggested that sometimes differences might "be large" (TNS-BMRB 2013, p.10). On some measures, the profile of the sample deteriorated when postal questionnaires were added, despite an increase in response rate.

[30] There were 1,146 respondents to the pilot survey, of whom 968 completed all the survey sections. For the second survey, 58 percent of respondents to the second survey were women, and 92% white. Fifty-three percent were in work, 25 percent retired, 6 percent looking after the home, 5 percent in full-time education, 5 percent unable to work because of a long-term illness or disability and one percent unemployed and looking for work. Seven per cent of respondents were aged between 16 and 24, 32 percent from 25 to 44, 36 percent from 45 to 64, 18 percent from 65 to 74 and 8 percent were 75 or older. Thirty-seven per cent owned their own home outright, 27 percent owned their home with the help of a mortgage, two percent had shared ownership and 27 percent were renting their home. Forty-one per cent had a degree of equivalent qualification, 44 percent another form of qualification and 15 percent no qualifications. Twenty-six per cent reported a long-term limiting illness or disability.

[31] In the pilot survey, 12,047 sampled addresses yielded 968 responses. In the second survey, 10,000 addresses yielded 872 complete responses. While this constituted a low response rate, the aim was to construct a scale of legal confidence, which has very different requirements to an exercise focussed on producing population estimates. The most important consideration for scale development is to ensure a broad range of perspectives, demographics and levels of confidence, rather than to minimise total survey error (Weisberg, 2005). More generally, numbers of respondents were more than adequate for scale development and Rasch analysis (Linacre 1994).

Drawing on the self-efficacy literature, various approaches can be taken to drafting questions for a general legal confidence scale.

A first is to simply ask about respondents' confidence in their ability to resolve justiciable problems, as Gramatikov and Porter (2011) did.

A second is to ask about respondents' confidence in relation to tasks and abilities relevant to resolving justiciable problems, as in Mackie's (2013:10-11) operationalisation of Collard et al.'s (2011) capability framework.

Our surveys included item pools adopting both of these approaches; but we do not expand upon them in this paper. The first approach did not prove particularly successful,[32] and while the second approach did yield scales, they did not go to general legal confidence and were less robust than the scale we detail below.

A third approach – that we expound upon in this paper – is to ask about respondents' confidence in their ability to resolve justiciable problems as law specific scenario elements become increasingly demanding. The approach is thus centred upon a set of questions centred on a scenario escalation. This approach recognises Bandura's (1997, p.42) concern that self-efficacy measures should "be tailored to domains of functioning and … represent gradations of task demands within those domains," and is well-illustrated by Vealey and Chase's (2008, p.75) examples of performance levels in the context of pitching in baseball:

> "How certain are you that you can throw a curveball for a strike in these situations?
>
> - In practice?
> - When warming up for a game?
> - In early innings with no school and no runners on base?
> - In middle endings with the score is tied and runners on base?
> - In the late innings with the score tied?
> - In late evenings with the score tied and runners on base?
> - In the final innings with a one run lead and runners on base?
> - In the final ending with a one run lead, do you out, full count, and bases-loaded?"

It was hoped that using a progressing scenario set in a legal context – with law scenario elements becoming increasingly demanding – would promote engagement (improving quality of responses and respondent fit) and improve the ability of a scale to differentiate between respondents with differing levels of confidence. The scenario we used and the twelve items were:

> If you found yourself facing a significant legal dispute – such as being unreasonably sacked by your employer, injured as a result of someone else's negligence, involved in a dispute over money as part of a divorce, or facing eviction from your home – how confident are you that you could achieve an outcome that is fair and you would be happy with in the following situations?
>
> a. Disagreement is substantial and tensions are running high.

[32] The first approach gave rise to a variety of concerns. The definition of even simple scenarios in our approach involved substantially increased administration time; thus limiting the number of questions it was feasible to ask. Moreover, this approach was beset by differential item functioning. Consequently, we were not able to identify a set of questions that was appropriate for use as a scale.

b. The other side says they 'will not rest until justice is done'.
c. The other side is represented by a solicitor, but you are not.
d. The other side refuses to speak to you except through their solicitor.
e. The other side threatens you with 'legal action'.
f. You receive a letter from the solicitor threatening court action.
g. You receive notice from a court stating that legal proceedings have been commenced against you.
h. The notice also says you must complete certain forms, including setting out your case.
i. You receive a letter telling you that you must appear in court.
j. The problem goes to court, a barrister represents the other side, and you are on your own.
k. As you present your case, the other side's barrister argues that much of your evidence is inadmissible or irrelevant.
l. The court makes a judgement against you, which you see as unfair. You are told you have a right to appeal.

Twelve items constitutes a relatively small item pool size,[33] reflecting the difficulty of generating items in the progressing scenario format.[34] Furthermore, the number was an expansion on an encouraging five-item scale produced using pilot survey data.

Our final measurement format took the form of a four-point Likert scale. In the pilot questionnaire, we tested a 1 to 100 response format. Using Rasch analysis, continuous items need some attention to fit the model well. The typical approach to this issue is to convert the continuous score into smaller discrete 'chunks', acknowledging that the continuous form is over-precise. Further guidance can be found in Linacre (2007, 2015). Converting the continuous scale into four categories appeared to work well during the pilot, so a four-point Likert scale was adopted (very confident, quite confident, not very confident, not confident at all). Evaluation of items and optimisation of the scale length was conducted using Rasch analysis, the conduct of which is described in detail below.

C. Analysis

Rasch analysis (e.g. Boone et al. 2014, Bond & Fox 2015) was used to develop a scale of general legal confidence; representing the first use of such methods scale development in an empirical legal context. For a unidimensional set of items (i.e. items that measure a single trait), Rasch analysis can be used develop and refine a scale. It allows detailed examination of the functioning of scale as a whole, how individual respondents and items fit, and can be used to develop a scoring protocol.

1. Software

Specialist software is typically required to conduct Rasch analysis, and this, as well as relative technical difficulty, have been cited as a barrier to wider use (e.g. Hays et al. 2000, Hobart & Cano 2009). Common software used to implement Rasch analysis include RUMM2030 (Andrich et al. 2016), used in the current analysis, and WinSteps (Linacre 2016).[35]

[33] It is not uncommon to begin with an item pool four times larger than the final scale, though where items are difficult to generate they may be as small as fifty percent larger than the final scale (DeVellis, 2012), which in our case would be nine items.
[34] As well as pressures on space within the questionnaire.
[35] For a complete list, see https://www.rasch.org/software.htm.

2. Choice of Rasch Model

There are two main types of Rasch model that may be used with polytomous data; the rating scale model and partial credit model. The partial credit model is the default within RUMM2030, placing no constraints on threshold parameters and allowing them to vary by item. In contrast, in the rating scale model, items share the same rating scale structure. A simple introduction to the two approaches can be found in Bond and Fox (2015), with a short description of advantages and disadvantages in Wright (1998) and Linacre (2000). One common approach (e.g. Persson et al., 2014; Vincent et al., 2015) to determine model choice is the use of a likelihood ratio test to test the efficiency of the parameterisation employed for the unrestricted form of the model (RUMM Laboratory, 2013a). The test, which is available within the RUMM2030 software, assesses the unrestricted parameterisation (partial credit model) against the rating re-parameterisation. A non-significant outcome indicates that the simpler rating scale model should be adopted (since no additional information is available in the unrestricted version, while a significant outcome indicates that the partial credit model should be used.[36]

3. Fit to the Rasch model

A number of measures of fit were considered. Overall fit was assessed using an item-trait interaction statistic. This is reported in RUMM2030 as a chi-squared statistic and should be non-significant (following a Bonferonni correction for the number of items in the scale). A significant value would indicate that hierarchical ordering of items varies across the trait (e.g. confidence) which would compromise the required property of invariance (Tennant & Conaghan 2007). Two item-person interaction statistics were also considered (for items and persons). In each case, fit is represented by a z-score, where perfect fit would have a mean of zero and standard deviation of one.[37] In practice, a fit residual standard deviation value of 1.5 or less (for items and persons) is commonly considered to indicate acceptable fit.

Individual item and person fit was also examined as both residuals as well as chi-squared statistics (and Bonferonni adjusted p-values). Items of concern were indicated by fit residuals below -2.5 or above 2.5.[38] Values below -2.5 typically indicate overfit or redundancy,[39] with the item a possible candidate for removal on this basis (i.e. it is already captured by other items, or there may be a violation of local (response) independence, which is discussed further below). Values above 2.5 indicate an underfit item.[40] This could be due to a number of reasons

[36] Further details on the derivation of the likelihood-ratio statistic can be found in RUMM Laboratory (2013a).

[37] For item fit, for each item, the statistic is based on the standardised residuals of the responses of all persons to the items. Residuals are squared and summed over persons, and transformed to make it more nearly approximate to a standard normal deviate. The hypothesis is that if the data (items) fit the model, then the deviations between the responses and the model are no more than random errors (i.e. with a mean close to zero and standard deviation near to one, but not exceeding 1.5). The person statistic is constructed in much the same way as for items (RUMM Laboratory, 2013b).

[38] This range is universally employed by Rasch papers using RUMM software (and recommended by RUMM Laboratory (2013b)). Calculation of individual item and person fit statistics in RUMM2030 are related to, but calculated differently from those in some other software (e.g. WINSTEPS). While none of the fit statistics are identical across packages, OUTFIT ZSTD in WINSTEPS and residual statistics in RUMM2030 are comparable, because of their standardized nature across all persons (Tennant and Conaghan, 2007).

[39] Observations of means in successive class intervals are steeper than the item characteristics curves.

[40] Observations of means in successive class intervals are flatter than the item characteristics curves.

(including possible violation of unidimensionality), with a range of additional diagnostics allowing further assessment, and possible correction of a misfitting item (see below). If an individual item fit cannot be improved, it may need to be removed.[41] As with items, misfitting persons were identified by fit residual values above 2.5. Such cases can seriously affect the fit at an item level (Tennant & Conaghan 2007). In such cases, response patterns were examined prior to removal (which for example, might show a respondent who was not fully engaged with the exercise). In the present study, the hope was that presenting a scenario with progressing items might help to engage respondents and reduce issues with misfitting respondents.

4. Internal consistency

Internal consistency of a scale was assessed using the Person Separation Index (PSI), which gives a measure of a scale's ability to discriminate between individuals with varying levels of the trait (e.g. confidence). Although acceptable levels may vary depending on the scale and its use, PSI can be interpreted in a similar fashion to Cronbach's Alpha (but it uses the person estimates in logits instead of the raw scores), with values exceeding 0.7 indicating acceptable internal consistency (Nunnally, 1978). Lower values may indicate the need for additional items. Person separation index can also be used to indicate how many statistically distinguishable measurement levels exist in the sample. For example, a person separation index of 0.61 allows two, 0.80 three and 0.88 four strata (Linacre, 2013). Further details on separation levels and strata, including how they are derived are set out in '9. Scoring a scale' below.

In this study the number of items that could be included was limited by the number of suitably distinct progressing items that could be created. The hope was that a pool of twelve items would be enough to yield a final scale with acceptable internal consistency.

5. Discrimination

Discrimination was also assessed graphically by examining the Item Characteristic Curve (ICC) for each item, with poor discrimination a common contributor to item misfit. The ICC allows examination of the extent to which an item deviates from the model. For polytomous items (e.g. the Likert type items in the current study), the ICC shows the expected response value for each possible location on the legal confidence continuum. The expected values on the y-axis range from 0 (not confident at all) to 3 (very confident). The location of an item corresponds to an expected value of 1.5.[42] Dots on figure illustrate the observed means of people placed into (in our case) nine adjacent class intervals, based on person ability measures.[43] If the data fit the model, the means of people in each class interval should be close to the curve. The curve together with the class interval means give a good indication of how well the data for an item conforms to the model. If mean values depart from the curve, this will often be accompanied by poor individual or overall item fit (as discussed above). Departures might include over-discriminating items (means for each class interval form a steeper pattern than the ICC) or under-discriminating item (means for each class interval are more shallow

[41] Changes following item deletion (or the model more generally) can be further assessed with a cross-validation sample (i.e. to test whether the improvement in fit following removal may be inflated). This might involve verifying the model with a subset of respondents, for example, by splitting the sample at the outset (De Jong-Gierveld and Kamphuis, 1985) or using a separate cross-validation sample (Kliem et al., 2015). In our case, we verified the scale with a separate cross-validation sample as part of a UK-wide face-to-face omnibus survey.

[42] If we were dealing with dichotomous items, it would correspond to a 50% chance of success.

[43] This involves sorting person location values and dividing then into class intervals with approximately equal numbers (RUMM Laboratory, 2013b). Number of class intervals can be altered prior during analysis in RUMM2030 to ensure reasonable numbers in each interval. In our case, numbers in class intervals varied from 64 to 105.

than the ICC) (RUMM Laboratory, 2013a). Remedial action may involve removing items with poor discrimination. Figure 1 in the results section illustrates ICCs for legal confidence two items.

6. Suitability of the Response format

Item responses took the form of polytomous (four-point) Likert scales. Suitability of the response format was checked by examination of the threshold map and category probability curves for individual items, which illustrates category structure. Where individuals respond in a manner consistent with their level of the trait, thresholds should be ordered. Disordered thresholds can occur, for example, where there are too many response options or respondents struggle to differentiate between options. Remedial action may include rescoring response categories (i.e. to fewer categories). This an important step if misfitting items are identified as a possible contributor to lack of fit. An example of ordered thresholds is provided by Figure 2 in the results section. As can be seen, Likert responses (from 0 to 3) are in order and take it in turns to have the highest probability of endorsement as the underlying trait (confidence) increases.

7. Response dependence

Response dependence was also tested as part of the Rasch analysis. Local independence is a requirement of the Rasch model and means that having extracted the Rasch factor (i.e. the confidence scale) there should be no leftover patterns in the residuals. Response dependence is an issue where response to one item depends on response to another item. A commonly cited example is that of numerous walking items in the same scale (Tennant & Conaghan 2007), where, if a person can walk a mile without difficulty, they must necessarily be able to walk a lesser distance without difficulty. Response dependence can inflate reliability (indicated by an artificially high PSI) and affect parameter estimates in Rasch analysis (Tennant and Conaghan, 2007). In practice, examination of the residual correlation matrix for values over 0.2 indicates potential response dependence and redundancy (Marais & Andrich 2008), with Table 3 in the results section illustrating a residual correlation matrix for the twelve legal confidence items. Where values exceed 0.2, the first step is to carefully check the wording of the items. Sub-test analysis, re-running Rasch analysis having combined dependent items, can be used to assess the extent to which reliability has been inflated (Marais, 2013). Remedial action may involve the rewording or removal of items.

This was a particular concern for our items since they took the form of a progressing scenario. The progression was designed to engage respondents and yield a range of item difficulties, in order to enhance scale targeting and ability to differentiate between respondents with different levels of confidence. However, the progression may also increase the chance of response dependence (since each item flows from the previous item), which would need to be carefully assessed.

8. Unidimensionality

Rasch analysis requires that items form a unidimensional scale.[44] Testing dimensionality as part of the Rasch analysis used the procedure set out by Smith (2002) and described in Tennant and Conaghan (2007). Following a principal components analysis (PCA) of the residuals,

[44] Though there have been multidimensional extensions of the unidimensional Rasch model. Foe example, the Multidimensional Random Coefficient Multinomial Logit (MRCML) model described by Briggs and Wilson (2003).

correlations between items and the first residual factors are used to define two subsets of items. An independent t-test is then used to test the difference in person estimates between the two subsets, with a non-significant result indicating no evidence of multidimensionality.

The progressing scenario items were defined to measure what was hoped would be a single trait (general legal confidence). If multiple traits are identified (i.e. the data is multidimensional), items need to be separated by trait and Rasch analysis conducted for separate traits.

9. Differential item functioning

Differential item functioning (DIF), which can have an impact on model fit, occurs when particular groups (e.g. men and women, younger and older respondents) perform differently on an item despite having comparable levels of the trait being measured. Graphically, DIF can be explored by superimposing groups of interest (e.g. younger and older respondents) on item characteristic curves to assess whether or not they perform differently. DIF is also assessed statistically using Analysis of Variance to compare scores for each level of the person factor (older, younger) and across levels of the trait (class intervals). There are two distinct forms of DIF (Teresi et al. 2000); Uniform DIF, where consistent systematic difference is observed in the groups' responses to an item, would be indicated by a significant person factor. Remedial action here might involve separate calibration of the item for each group, though this has the disadvantage of adding complexity to the scoring of a scale. Non-uniform DIF, where differences vary across levels of the trait, would be indicated by a significant person factor by class interval interaction. In this case, it is likely that the item would need to be removed.

10. Scale targeting

Scale targeting was assessed graphically through examination of the person-item distribution, which illustrates individuals' scores and item placement on the underlying trait (e.g. confidence, expressed in logits). In a well targeted scale, items would span the full range of individual scores. This indicates that a scale is not too easy (e.g. items that nearly all respondents would be confident about) and not too hard (e.g. items that nearly all respondents would not be confident about). Given that the scale is centred on zero logits, targeting can also be examined by how close the mean location value for persons is to zero. A poorly targeted scale is likely to indicate a need for additional easier or harder items to fully span individuals' scores or replacement of items with similar difficulty to give a broader spread of difficulty.

In the present study it was hoped that the progressing items would yield items of different (and generally increasing) difficulty, which would hopefully help to span a wide range of respondent scores, and help to differentiate between respondents with a relatively modest number of items.[45]

11. Scoring a scale

Once a scale is developed, satisfying the various requirements/diagnostics set out above, it can be used to produce a score. This involves providing guidance on how to calculate raw scores from responses, and a conversion table to change raw scores into Rasch converted. As discussed previously, the Rasch location values have been converted from an ordinal to an

[45] It is also worth noting that items of variable difficulty (e.g. in our progressing scenario) are better suited to Rasch methods rather than CTT. As discussed above CTTs psychometric properties are at an overall test level rather than at an individual item level. Rasch models psychometric properties are at the item, rather than the test level, taking advantage of the progressing scenario design and variable difficulty.

interval scale (Wright & Linacre 1989), making them appropriate for a wider range of common statistical analyses. For ease of use, Rasch location values are also commonly converted to alternative ranges (e.g. 0-100).

A further useful way to interpret the person separation index is to convert it into strata or separation levels (Fisher, 1992; Linacre, 2013; Wright & Masters, 2002) which indicate how many statistically distinguishable measurement levels exist in the sample. Strata are considered a refinement of separation (Linacre, 2013) where high and low measures are considered valid levels of performance, which seems a reasonable assumption in the our context. Equation 2 and 3 illustrate how separation levels are derived from person separation index (PSI) and how strata are derived from separation levels.

Equation 2. Deriving separation levels from person separation index

$$Separation\ (G) = \sqrt{\frac{PSI}{(1 - PSI)}}$$

Equation 3. Deriving strata from separation levels

$$Strata = \frac{(4G + 1)}{3}$$

So for example, a person separation index of 0.8 would allow two levels of separation and three strata (e.g. low, medium and high). Linacre (2013) also sets out the approximate percentage of samples in each separation or strata (for approximately normally distributed samples).[46] For examples of the use of strata with Rasch analysis see de Haan et al., (2011), Prieto et al., (2003) and Duncan et al., (2003).

12. Construct Validity

Construct validity (or external validity) concerns "the degree to which measures are related to external measures of the same construct, similar constructs, and other constructs" (Wolfe & Smith, 2007: 268).[47] In our study, no external measures of the same 'legal confidence' construct exist. However, we did include a number of questions designed to measure similar, related constructs, allowing examination of construct validity by assessing the relationship between legal confidence scores and responses to these variables. Specifically, we collected information on (non-criminal) legal problem experience in the past five years,[48] whether respondents felt problems had been handled well or not and whether or not they felt problem outcomes were fair. Data were also collected on use of lawyers in the past five years, and whether or not respondents were satisfied with the help provided. It is hypothesised that legal confidence is higher where problems are well handled, had positive outcomes and where respondents were

[46] For two strata, 50 percent at each level; for three, 23, 54 and 23 percent; for four, 14, 36, 36 and 14 percent. See Linacre (2013) for further details.

[47] This aspect of validity is often overlooked in scale development. Wolfe and Smith (2007) also set out a comprehensive review of aspects of validity in Rasch instrument development and related forms of evidence or analysis.

[48] Including disputes with an ex-partners over arrangements for children, disputes with employers, disputes with landlords, falling behind on rent or mortgage payments, consumer issues concerning faulty goods, significant injuries that were somebody else's fault and disputes over a will.

satisfied with advice from lawyers, and lower where respondents felt problems were not well handled, felt the outcome was negative or were dissatisfied with advice from lawyers.[49]

III. RESULTS

A. Rasch analysis

Responses for each item are set out in Table 1.

Table 1: Responses to the twelve scenario progression questions[50]

	Very confident		*Quite confident*		*Not very confident*		*Not confident at all*	
	N	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Item a	35	4.2%	414	49.2%	359	42.6%	34	4.0%
Item b	48	5.7%	438	51.8%	320	37.8%	40	4.7%
Item c	16	1.8%	136	15.7%	470	54.3%	243	28.1%
Item d	41	4.8%	358	41.8%	364	42.5%	94	11.0%
Item e	48	5.6%	353	41.0%	343	39.9%	116	13.5%
Item f	42	4.9%	321	37.3%	357	41.5%	140	16.3%
Item g	26	3.0%	290	33.8%	386	45.0%	156	18.2%
Item h	72	8.3%	432	49.7%	274	31.5%	91	10.5%
Item i	66	7.6%	384	43.9%	303	34.7%	121	13.8%
Item j	24	2.7%	111	12.7%	392	44.9%	347	39.7%
Item k	17	2.0%	136	15.8%	399	46.4%	308	35.8%
Item l	39	4.5%	234	27.0%	395	45.5%	200	23.0%

According to the likelihood-ratio test the general legal confidence items did not meet the requirements of a rating scale model, with a partial credit model implemented.[51] Rasch analysis was undertaken on the responses of 761 respondents who answered all twelve legal confidence items.[52] Including all twelve items, a significant item-trait interaction (overall fit) suggested deviation from the model ($\chi^2_{108} = 342.39$, $p < 0.001$[53]). A fit residual standard deviation of 3.40 for items also indicated significant misfit, with a value of 1.49 for respondents just below the cut-off for acceptable fit (of 1.5). A fit residual mean value of -0.68 suggested a reasonably well targeted scale (which was confirmed by examination of the person-item location distribution),[54] while a person separation index of 0.93 was high, indicating good internal consistency and ability to discriminate between respondents with differing levels of confidence. Nonetheless, misfitting items in particular needed to be addressed. Item fit for all twelve items is shown in Table 2. As shown, there were a number of items with fit residuals

[49] Court et al., (2010) adopt a similar approach, assessing construct validity by the extent to which a scale has the expected relationship with other related variables. In their case, scores for a six-item anxiety scale were compared between patients attending general practitioners for routine or emergency appointments. They hypothesized that patients attending for an emergency appointment would report significantly higher levels of anxiety.

[50] Numbers of 'don't know' responses were relatively small, with 59 for item a, 55 for b, 36 for c, 44 for d, 41 for e, 41 for f, 43 for g, 32 for h, 27 for i, 27 for j, 41 for k and 33 for l.

[51] A significant likelihood-ratio test indicated that the partial credit model should be adopted; $\chi^2_{21} = 205.40$, $p < 0.001$.

[52] Of 888 respondents who answered at least one of the twelve items.

[53] Which is less than the Bonferroni corrected p-value of 0.004 (0.05/12 on account of the twelve items in the scale).

[54] A particularly well targeted measure should have a value around zero, with very high or low values indicating generally high or low scores and possibly a scale that is not too well suited to the cohort.

greater than 2.5 (items a and l) or less than -2.5 (items i, f, g and l, indicating possible overfit/redundancy).

Table 2: Fit of all Twelve Legal Confidence Items to the Rasch Model

Item	*Location*	*SE*	*Fit residual*	*DF*	*χ2*	*DF*	*p*
Item a	-0.914	0.077	2.72	685.5	41.589	9	< 0.001
Item b	-1.106	0.074	1.258	685.5	24.927	9	0.003
Item c	1.414	0.074	-0.997	685.5	25.929	9	0.002
Item d	-0.418	0.07	0.271	685.5	28.054	9	< 0.001
Item e	-0.439	0.068	-5.943	685.5	27.419	9	0.001
Item f	-0.243	0.068	-6.694	685.5	24.487	9	0.004
Item g	0.297	0.071	-6.106	685.5	27.854	9	0.001
Item h	-0.964	0.068	0.481	685.5	20.365	9	0.016
Item i	-0.644	0.067	-3.687	685.5	22.459	9	0.008
Item j	1.34	0.07	-2.43	685.5	39.534	9	< 0.001
Item k	1.396	0.071	-1.81	685.5	44.837	9	< 0.001
Item l	0.283	0.067	3.099	685.5	14.936	9	0.093

Note: a significant p-value would be less than a Bonferroni adjusted value of 0.05/12 = 0.00083, with p-values for items a, j and k significant.

Further exploration of items showed no evidence of disordered thresholds and principal components analysis of the residuals confirmed the items were unidimensional. Importantly, however, examination of the residual correlation matrix suggested response dependence between items as illustrated in Table 3. Of particular concern were the high correlation values between items e and f, f and g and j and k.

Table 3. Person-item residual correlation matrix to examine response dependence

Item	Item a	Item b	Item c	Item d	Item e	Item f	Item g	Item h	Item i	Item j	Item k	Item l
Item a	1											
Item b	0.285	1										
Item c	-0.005	-0.041	1									
Item d	-0.075	0.02	-0.077	1								
Item e	-0.141	-0.106	-0.181	0.173	1							
Item f	-0.211	-0.155	-0.23	-0.009	0.405	1						
Item g	-0.227	-0.213	-0.223	-0.105	0.141	0.404	1					
Item h	-0.128	-0.114	-0.281	-0.139	-0.131	-0.071	-0.017	1				
Item i	-0.164	-0.185	-0.267	-0.25	-0.147	-0.072	0.042	0.254	1			
Item j	-0.226	-0.309	0.28	-0.238	-0.273	-0.277	-0.146	-0.13	0	1		
Item k	-0.204	-0.211	0.161	-0.14	-0.232	-0.251	-0.165	-0.23	-0.097	0.388	1	
Item l	-0.05	-0.074	-0.124	-0.175	-0.237	-0.224	-0.213	-0.047	-0.05	-0.04	0	1

Removing items f and k addressed response dependence, with no remaining residual correlations exceeding 0.2. However, there were still item fit residuals well outside standard acceptable limits. In particular, item e had a fit residual of -5.32 and item g a fit residual of -5.10, again indicating overfit/redundancy (since they were less than -2.5). Moreover, a significant item-trait interaction statistic ($\chi^2_{80} = 218.93$, $p < 0.001$) and item fit residual standard deviation value exceeding 1.5 (2.75) indicated lack of fit to the model. Having removed item e, large negative fit residuals of -4.42 and -4.44 remained for items g and i respectively, with a significant item-trait interaction statistic ($\chi^2_{72} = 188.36$, $p < 0.001$) and item fit residual standard deviation of 2.35 still indicating lack of fit. A significant item-trait interaction statistic ($\chi^2_{64} = 131.31$, $p < 0.001$) and item fit residual standard deviation value exceeding 1.5 (1.84) remained having removed item i, as well as a negative fit residual for item g of -3.86. Removing item g yielded a significant item-trait interaction statistic ($\chi^2_{56} = 92.56$, $p = 0.002$) and item fit residual standard deviation slightly above acceptable limits at 1.55. All items had fit residuals between -2.5 and 2.5 with the exception of item c (-3.19).

Removing item c resulted in a final six-item scale (comprised of items a, b, d, h, j and l) produced a non-significant item trait interaction ($\chi^2_{48} = 60.08$, $p = 0.11$[55]), good item fit (fit residual standard deviation = 0.70) and good person fit (fit residual standard deviation = 1.15). Looking at items individually showed no further evidence of misfitting items, as illustrated in Table 3.

Table 3: Fit of the Remaining Six Legal Confidence Items to the Rasch Model

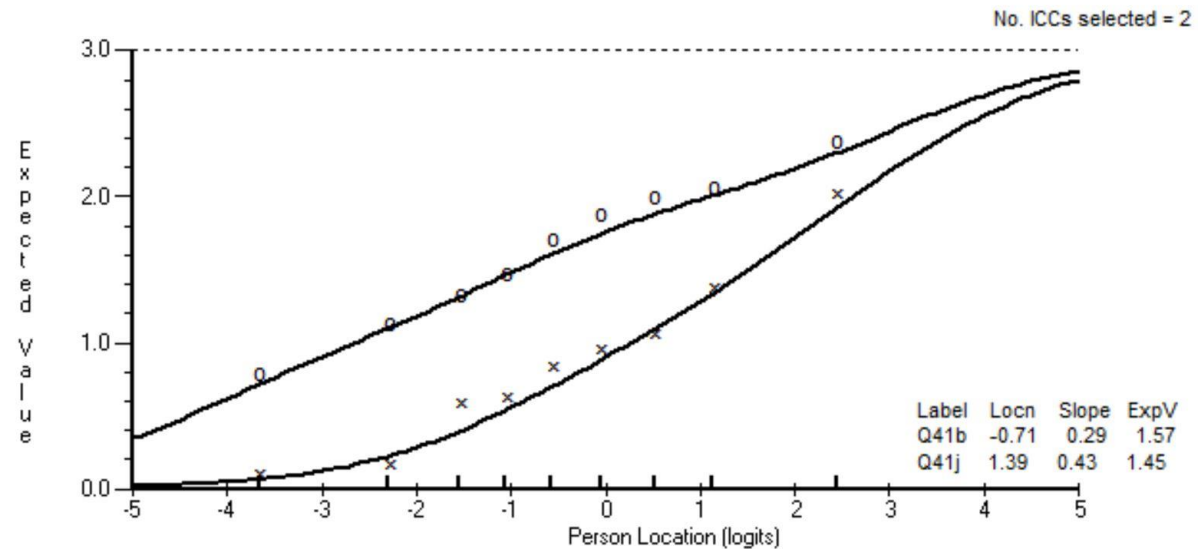
Item	*Location*	*SE*	*Fit residual*	*DF*	*χ2*	*DF*	*p*
Item a	-0.574	0.075	0.090	623	12.767	8	0.120
Item b	-0.711	0.071	-1.843	623	5.156	8	0.741
Item d	-0.065	0.066	-0.411	623	5.589	8	0.693
Item h	-0.569	0.064	-0.872	623	6.254	8	0.619
Item j	1.386	0.065	-1.258	623	20.259	8	0.009
Item l	0.534	0.063	-0.318	623	10.053	8	0.261

Note: a significant p-value would be less than a Bonferroni adjusted value of 0.05/6 = 0.0083.

A person separation index of 0.83 exceeded common cut-offs and suggested good internal consistency and ability to discriminate between respondents with differing levels of confidence. Since the person separation index exceeded 0.80 this also allows the scale to be split into three strata (Linacre, 2013). The item characteristic curves (for individual items) also indicated good discrimination and fit to the model (as illustrated for items b and j in Figure 1). As shown, the means of people in each class interval (illustrated by the dots) lay close to the curve for both items, indicating that the data for both items conformed to the model. The figure also illustrates the variation in difficulty between the items (with item b 'easier' than item j).

Figure 1: Item characteristic curve for item b and j.

[55] Which is greater than the Bonferroni corrected p-value of 0.008 (0.05/6 on account of the six items in the scale).



All thresholds were ordered (as illustrated in Figure 2 for item b), demonstrating the suitability of the response format. Ordered thresholds for all six items, as well as variation in the difficulty of items can be seen in Figure 3.

Figure 2: Category probability curves for item b

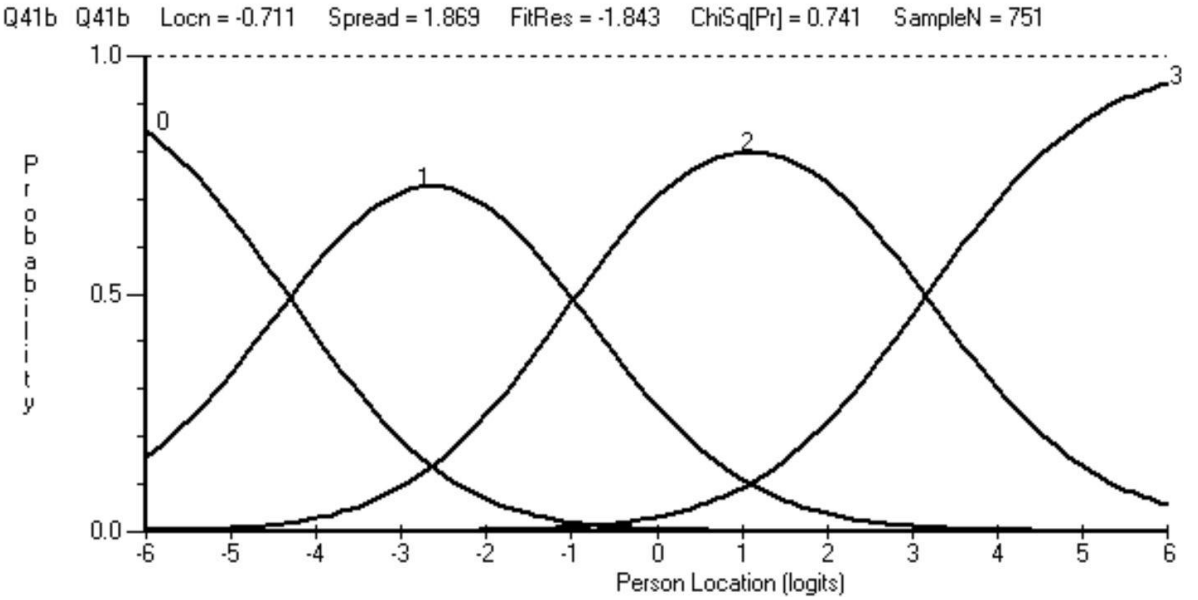
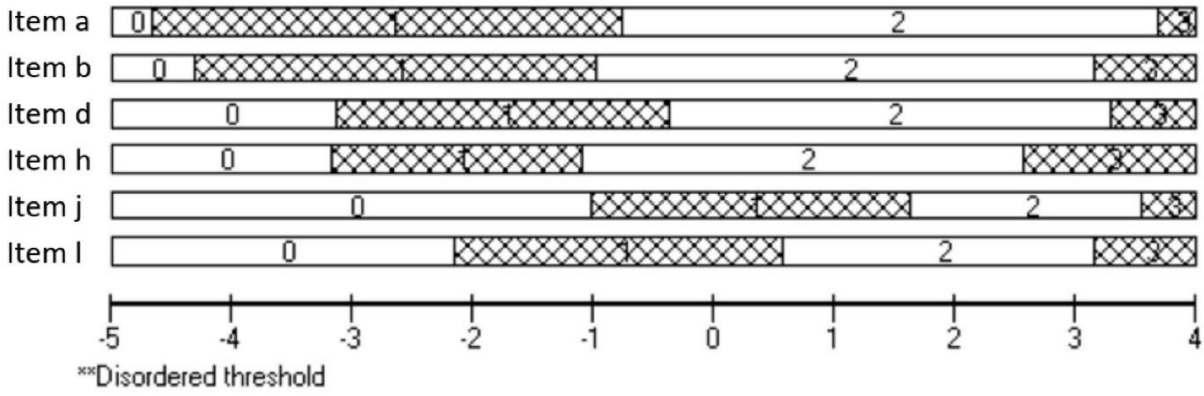
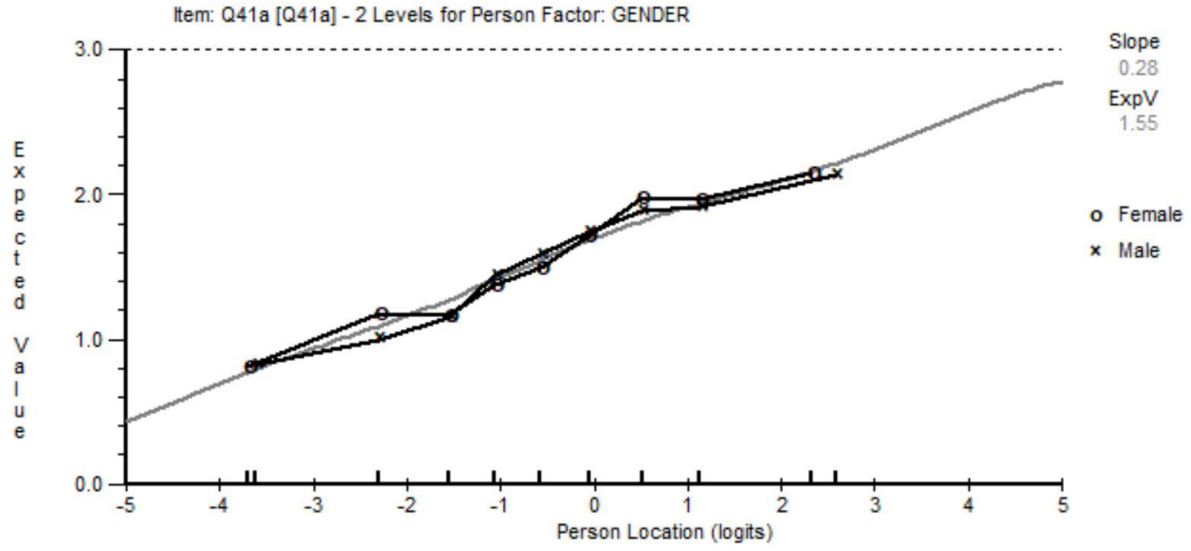


Figure 3: Threshold map for all six items in the final General Legal Confidence (GLC) scale



A final testing of dimensionality (PCA of the residuals) confirmed the items were unidimensional and there was no evidence of differential item functioning on the basis of gender, age, prior experience of legal problems or prior contact with lawyers or courts. Figure 4 provides an illustration of item characteristic curves for item a, split by gender. As can be seen, there is no evidence of differential item functioning, which would be indicated by separation between males and females.

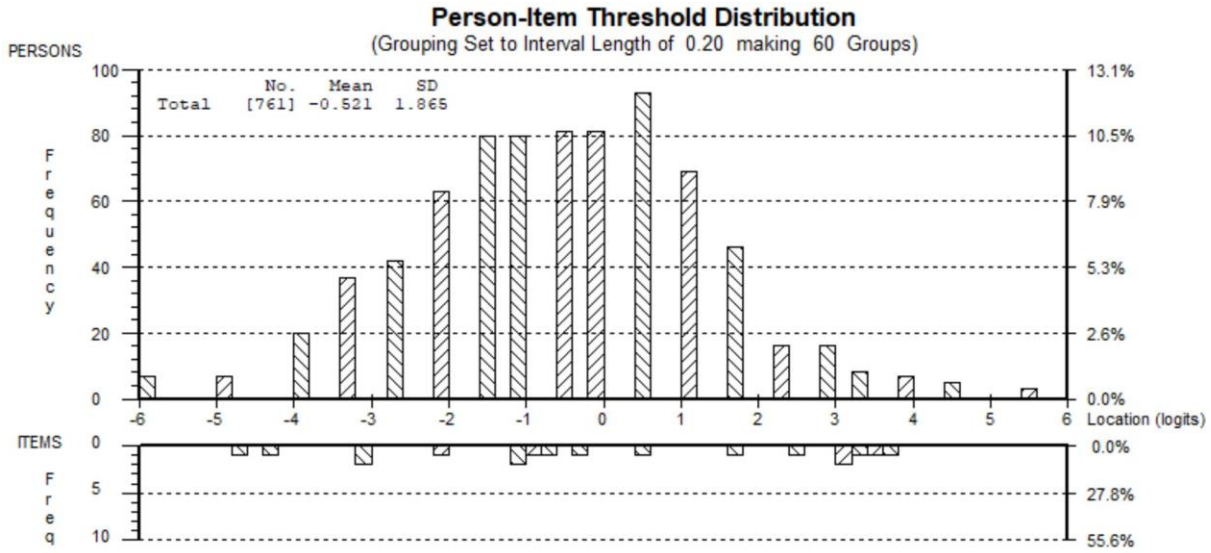
Figure 4: Item characteristic curves for item a, split by gender.



Finally, examination of the person-item distribution indicated that the scale was well targeted. A mean of -0.52 (relatively close to zero) suggested the scale was not too difficult or too hard, while the items spanned a good range of respondents (i.e. they varied in difficulty) with a relatively small number of items (Figure 5).[56]

Figure 5: The person-item distribution for the final six-item General Legal Confidence GLC scale.

[56] Item thresholds of the final six items ranged -4.1 logits to 4.3 logits, spanning a fairly full range of the trait (as indicated in Figure 5). Gaps between adjacent thresholds were also in acceptable ranges, varying from 1.9 to 4.4 logits (i.e. at least 1.4 logits to show empirical distinction, but not more than 5 to avoid large gaps in the variable (Linacre, 1999)).



A cross-validation sample of respondents to a UK-wide face-to-face omnibus survey[57] confirmed the psychometric properties of the six-tem GLC scale. Both item fit and person fit were acceptable, while a person separation index of 0.89 exceeded the value of the initial sample, again indicating good internal consistency. All item thresholds were ordered, confirming the suitability of the response format. There was no evidence of differential item functioning (on the basis of age or gender), response dependence or multidimensionality and the scale remained well targeted.

B. The General Legal Confidence (GLC) Scale

The final six-item 'General Legal Confidence' (GLC) scale was as follows (relabelling items a, b, d, h, j and l from 1-6);

> *If you found yourself facing a significant legal dispute – such as being unreasonably sacked by your employer, injured as a result of someone else's negligence, involved in a dispute over money as part of a divorce, or facing eviction from your home – how confident are you that you could achieve an outcome that is fair and you would be happy with in the following situations?*
>
> 1. *Disagreement is substantial and tensions are running high.*
> 2. *The other side says they 'will not rest until justice is done'.*
> 3. *The other side refuses to speak to you except through their solicitor.*
> 4. *A notice from court says you must complete certain forms, including setting out your case.*

[57] The IPSOS face-to-face omnibus (Capibus) survey is conducted on a continuous basis using a random location approach, where interviews fulfil a number of interviews within randomly sampled small geographic areas. While the approach constitutes a probability sample, it is not as methodologically robust (e.g. if the interest is in generalisation of findings) as many carefully conducted bespoke surveys. For example, response rates are not formally calculated and there is no scope for interviewer briefing. Nonetheless, it is a more defensible than widely used opt-in panel surveys and entirely appropriate as a means of providing a cross-validation sample in scale development. The cross-validation sample utilised 278 respondents who answered all GLC items, which exceeds the number required for 99 per cent confidence that no item calibration is more than half a logit away from its stable value (Linacre, 1994).

5. *The problem goes to court, a barrister represents the other side, and you are on your own.*
6. *The court makes a judgement against you, which you see as unfair. You are told you have a right to appeal.*

Responses take the form of a four-point Likert scale: very confident; quite confident; not very confident; not confident at all. [58]

C. Scoring the GLC Scale

After administration, responses are scored to yield, first, a 'raw' score, then a Rasch converted 'GLC score'. To calculate the raw score, responses of 'very confident' are assigned a score of 3, 'quite confident' a score of 2, 'not very confident' a score of 1 and 'not confident at all' a score of 0. Across the six items this yields individual scores of between 0 and 18. These scores are converted into GLC scores (ranging from 0 to 100) using Table 4. This uses location values from the Rasch analysis to converting raw scores to an interval scale. A higher score indicates greater legal confidence.

Table 4: Scoring for the Six-Item 'General Legal Confidence' (GLC) Scale

Raw score	Rasch location values	Rasch converted 'GLC' score
18	5.48	100
17	4.60	92.3
16	3.89	86.1
15	3.34	81.2
14	2.82	76.7
13	2.30	72.1
12	1.74	67.2
11	1.13	61.9
10	0.51	56.5
9	-0.05	51.5
8	-0.57	47
7	-1.06	42.7
6	-1.55	38.4
5	-2.02	34.3
4	-2.64	28.9
3	-3.27	23.3
2	-3.99	17
1	-4.80	9.4
0	-5.94	0

D. Initial Baseline Scores and Strata

The mean GLC score among 785 survey respondents was 47.5 (standard deviation = 16.3). The median score was 47.0 (25^{th} percentile = 38.4, 75^{th} percentile = 56.5). The minimum score was

[58] To enable the item to be understood and the scenario to advance smoothly, the wording of item 4 had to be slightly changed from the version used in the second survey: "The notice also says you must complete certain forms, including setting out your case."

0 and the maximum 100. A person separation index of 0.83 indicated that three strata (low, medium and high confidence) could be discerned. Ranges of 0-36 (low confidence), 37-58 (medium confidence) and 59-100 (high confidence) returned percentages as close to Linacre's (2013) guidance on the approximate percentage falling into three strata (as discussed above). Using these strata, 22.9 percent of the respondents had low confidence, 54.5 percent medium confidence and 22.5 percent high confidence. It should be noted, that while the survey provided an excellent sample for scale development, numbers were relatively small for baseline measurement, and subject to change if conducted with a larger scale.

E. Construct Validity

Table 5 shows mean GLC score by legal problem experience and (for those reporting legal problems) how well respondents felt problems were handled, legal problem experience and (for those reporting legal problems) whether respondents felt problem outcomes were fair and lawyer use and (for those using lawyers) satisfaction with lawyer use.

Table 5. Mean GLC score for problem experience and lawyer use groups

		GLC Score	
Questions	Experience	Mean	N
Legal problem experience and how respondents felt problems were handled (in the past 5 years)	No legal problem	47.04	547
	All quite/very well	51.23	190
	At least one not/not at all well	38.14	48
Legal problem experience and whether respondents felt problem outcomes were fair (in the past 5 years)	No legal problem	47.04	547
	All quite/very fair	52.66	144
	At least one not/not at all fair	42.15	93
Lawyer use and satisfaction with lawyer use (in the past 5 years)	No lawyer use	46.90	578
	Lawyer use - satisfied	51.56	164
	Lawyer use - not satisfied	39.76	42

Compared to those not reporting any legal problems in the past five years, having problems that respondents felt were well handled was associated with a 4.2 point increase in GLC score ($\chi^2_1 = 9.78$, $p = 0.002$). Conversely, where respondents reported one or more problem that they felt was not handled well, this was associated with an 8.9 point decrease in GLC score ($\chi^2_1 = 13.55$, $p < 0.001$). Similarly, compared to those not reporting any legal problems in the past five years, having problems that respondents felt resulted in fair outcomes were associated with a 5.6 point increase in GLC score ($\chi^2_1 = 14.02$, $p < 0.001$). In contrast, where respondents reported one or more problem that they felt resulted in an unfair outcome, this was associated with a 4.9 point decrease in GLC score ($\chi^2_1 = 7.42$, $p = 0.006$). Compared to those who had not used lawyers in the past five years, lawyer use that respondents were satisfied with was associated with a 4.7 point increase in GLC score ($\chi^2_1 = 10.72$, $p = 0.001$). Conversely, lawyer use that respondents were not satisfied with was associated with a 7.1 point decrease in GLC score ($\chi^2_1 = 7.73$, $p = 0.005$).

IV. CONCLUSION

Our study set out to develop a scale of legal confidence using established approaches to scale development (DeVellis 2012), modern psychometric methods and specifically Rasch analysis (Boone et al. 2014, Bond & Fox 2015). To date, much of the focus on legal capability has been in identifying capability domains (Parle 2009; Collard et al 2011; Pleasence et al. 2014), and

while legal confidence/capability measures have been integrated into studies (e.g. Gramatikov & Porter 2011) questions/scales to date have been developed in an ad-hoc fashion, with little reference to the scale development literature or attempt to test the psychometric properties of scales (using either classical test theory or modern psychometric methods). This has been a shortcoming in empirical legal studies.

In summary, we generated a legal confidence item pool made up of items representing a progressing scenario set in a legal context. A pilot survey was used to test the approach and further develop the pool and response format, resulting in a pool of twelve items with a four-point Likert format in a second survey. Rasch analysis was used to evaluate the items and optimise scale length. The initial Rasch model using all twelve items in our progressing scenario item pool indicated significant evidence of misfitting items. This was a result of overfit/redundancy and response dependence. rather than items failing to measure the same legal confidence trait. Essentially some items were not needed or were dependent upon each other, reflecting the challenge of producing items that are suitably different from/not dependent upon each other in a progressing scenario. Nonetheless, presentation of items as a progressing scenario rather than a set of discrete items was an important innovation in the study. One initial concern was that such a progression could result in dependence (Marais & Andrich 2008), artificially inflating reliability. While this proved to be the case, removal of items yielded a scale with good psychometric properties. Moreover, the progressing scenario presentation had some clear advantages in engaging respondents and ensuring a spread of item difficulty. This increased the ability to discriminate between respondents with different levels of confidence, despite a relatively small number of items.

Reduction of the twelve items to a final set of six resulted in a scale with good psychometric properties – the General Legal Confidence (GLC) Scale. It showed good overall fit, item fit, person fit, targeting (not too easy or difficult) and internal consistency (ability to discriminate between individuals). All items had ordered thresholds (respondents were able to differentiate between the four Likert descriptors), there was no response dependence, items were unidimensional and there was no evidence of differential item functioning (on the basis of gender, age, problem experience or legal experience).

With limits to legal aid (including substantial reductions in jurisdictions such as England and Wales), a refocusing of policy on legal service users' needs (Legal Services Commission 2006; Commonwealth Attorney-General's Access to Justice Taskforce 2009; Canadian National Action Committee on Access to Justice in Civil and Family Matters 2013) and new forms of service delivery (Smith, 2014) placing new demands placed on users, the ability to quantify legal confidence and capability is increasingly important. The GLC scale demonstrates that it is possible to arrive at robust and coherent law specific measures of aspects of legal capability – in this case, legal capability – through the careful design of questions and application of the modern psychometric scale development techniques. More generally, our study represents the first application of modern psychometric approaches to scale development in the empirical legal field, and provides a useful guide for those wishing to develop scales in the field in the future.

REFERENCES

Allen M. J. & W. M. Yen (2002) *Introduction to measurement theory*. Long Grove, Il: Waveland Press.

Andrich, D. (2004) "Controversy and the Rasch model: A Characteristic of Incompatible Paradigms?" 42 *Medical Care* 17.

Andrich, D. B. Sheridan & G. Luo (2016) RUMM2030 [computer software] Perth: Rumm Laboratories.

Balmer, N. J., A. Buck, A. Patel, C. Denvir & P. Pleasence (2010) *Knowledge, Capability and the Experience of Rights Problems*. London: PLENET.

Bandura, A. (1997) *Self-Efficacy: The Exercise of Control.* New York, NY: WH Freeman and Co.

Beattie, S., L. Hardy, J. Savage, T. Woodman & N. Callow (2011) "Development and Validation of a Trait Measure of Robustness of Self-Confidence," 12 *Psychology of Sport and Exercise* 184.

Bond, T. & C. M. Fox (2015) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3d ed. New York: Routledge.

Blanchin, M., J-B. Hardouin, T. Le Neel, G. Kubis, C. Blanchard, E. Mirallie & V. Sebille (2010) "Comparison of CTT and Rasch-Based Approaches for the Analysis of Longitudinal Patient Reported Outcomes," 30 *Statistics and Medicine,* 825.

Boone, W. J., J. R. Staver & M.S. Yale (2014) *Rasch Analysis in the Human Sciences*. Dordrecht: Springer.

Briggs, D.C. & M. Wilson (2003) "An Introduction to Multidimensional Measurement Using Rasch Models," 4 *Journal of Applied Measurement* 87.

Cameron, W.B. (1963) *Informal Sociology: A Casual Introduction to Sociological Thinking.* New York: Random House.

Canadian National Action Committee on Access to Justice in Civil and Family Matters (2013) *Responding Early, Responding Well: Access to Justice Through the Early Resolution Services Sector.* Ottawa: National Action Committee on Access to Justice in Civil and Family Matters.

Cella, D & C.H. Chang (2000) "A Discussion of Item Response Theory and its Application in Health Status Assessments," 38 *Medical Care,* S66.

Christensen, K. B., S. Kreiner & M. Mesbah (eds) (2013) *Rasch Models in Health*. Hoboken, NJ: Wiley.

Collard, S., C. Deeming, L. Wintersteiger, M. Jones & J. Seargeant (2011) *Public Legal Education Evaluation Framework.* Bristol: University of Bristol Personal Finance Research Centre.

Commonwealth Attorney-General's Access to Justice Taskforce (2009) *A Strategic Framework for Access to Justice in the Federal Civil Justice System*. Canberra: Attorney-General's Department.

Coumarelos, C., D. Macourt, J. People, H. M. McDonald, Z. Wei, R. Iriana & S. Ramsey (2012) *Legal Australia-Wide Survey: Legal Need in Australia.* Sydney: Law and Justice Foundation of New South Wales.

Court, H., K. Greenland & T.H. Margrain (2010) "Measuring Patient Anxiety in Primary Care: Rasch Analysis of the 6-item Spielberger State Anxiety Scale," 13(6) *Value in Health* 813.

Dawes, M. E., J. J. Horan & G. Hackett (2000) "Experimental Evaluation of Self-Efficacy Treatment on Technical/Scientific Career Outcomes," 28(1) *British Journal of Guidance and Counselling* 87.

De Haan, J., N. Schep, W. Tuinebreijer, P. Patka & D. den Hartog (2011) "Rasch Analysis of the Dutch Version of the Oxford Elbow Score," 2 *Patient Related Outcome Measures* 145.

De Jong-Gierveld, J & F. Kamphuis (1985) "The Development of a Rasch-Type Loneliness Scale," 9(3) *Applied Psychological Measurement* 289.

DeVellis, R.F. (2012) *Scale Development: Theory and Applications*, 3d ed. Thousand Oaks: Sage.

DeVellis, R.F. (2006) "Classical Test Theory," 44 *Medical Care,* S50.

Duncan, P. W., R. K. Bode, S. Min Lai & S. Perera (2003) "Rasch Analysis of a New Stroke-

Specific Outcome Scale: the Stroke Impact Scale," 84(7) *Archives of Physical Medicine and Rehabilitation,* 950.
Engelhard, G. (2013) *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences.* New York: Routledge.
Fan, X. (1998) "Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics," 58 *Educational and Psychological Measurement,* 357
Felstiner, W. L. F., R. L. Abel & A. Sarat (1981) "The Emergence and Transformation of Disputes: Naming, Blaming, Claiming …" 15(3-4) *Law and Society Review* 631.
Fisher, W. (1992) "Reliability, Separation, Strata Statistics," 6(3) *Rasch Measurement Transactions* 238.
Galanter, M. (1974) "Why the "Haves" Come out Ahead: Speculations on the Limits of Legal Change," 9(1) *Law & Society Review* 95.
Galanter, M. (1976) "The Duty Not to Deliver Legal Services," 30 *University of Miami Law Review* 929.
Genn, H. (1999) *Paths to Justice.* Oxford: Hart.
Goldberg, D. & P. Williams (1988) *A User's Guide to the General Health Questionnaire*. Windsor: NFER-Nelson.
Goldstein, H. (1979) "The Mystification of Assessment," 22(1) *Forum for the Discussion of New Trends in Education* 14.
Goldstein, H. (2010) *Rasch Measurement: A Response to Payanides [sic.], Robinson and Tymms.* Available at www.bristol.ac.uk/cmm/hg/response-to-panayides.pdf.
Goldstein, H. (2015) "Rasch Measurement: A Response to Payanides, Robinson and Tymms," 41(1) *British Educational Research Journal*, 176.
Golub S, McQuay K (2001) Law and Policy Reform at the Asian Development Bank. Manila: Asian Development Bank.
Gramatikov, M.A. & R. B. Porter (2011) "Yes I Can: Subjective Legal Empowerment," 18(2) *Georgetown Journal on Poverty Law and Policy* 169.
Grembowski, D., D. Patrick, P. Diehr, M. Durham, S. Beresford, E. Kay & J. Hecht (1993) "Self-Efficacy and Health Behavior Among Older Adults," 34(2) *Journal of Health and Social Behavior* 89.
Habermas, J. (1987) *The Theory of Communicative Action, Volume 2: Lifeworld and System.* Cambridge: Polity Press.
Hays, R. D., L. S. Morales & S. P. Reise (2000) "Item Response Theory and Health Outcome Measurement in the 21st Century," 38 *Medical Care* 1128.
Hays, R. D. & J. Lipscomb (2007) "Next Steps for use of Item Response Theory in the Assessment of Health Outcomes," 16 (Supplement 1) *Quality of Life Research* 195.
HiiL (2014) *Justice Needs in Indonesia 2014: Problems, Processes and Fairness.* Den Haag: HiiL.
Hobart, J. & S. Cano (2009) "Improving the Evaluation of Therapeutic Interventions in Multiple Schlerosis: The Role of New Psychometric Methods," 13(12) *Health Technology Assessment* 1.
John, O. P., L. P. Naumann & C. J. Soto (2008) "Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues," in O. P. John, R. W. Robins & L. A. Pervin, eds., *Handbook of Personality: Theory and Research.* New York, NY: Guilford Press.
Kempson, E., S. Collard & N. Moore (2005) *Measuring Financial Capability: An Exploratory Study.* London: FSA.
Kliem, S., J. Beller, C. Kroger, Y. Stobel-Richter, K. Hahlweg & E. Brahler (2015) "A Rasch Re-Analysis of the Partnership Questionnaire," 5(2) *Sage Open* 1.

Legal Services Commission (2006) *Making Legal Rights a Reality: The Legal Services Commission's Strategy for the Community Legal Service*. London: Legal Services Commission.

Linacre, J. M. (1994) "Sample Size and Item Calibration Stability," 7(4) *Rasch Measurement Transactions* 328.

Linacre, J.M. (1999) "Investigating Rating Scale Category Utility," 3(2) *Journal of Outcome Measurement* 103.

Linacre, J. M. (2000) "Comparing 'Partial Credit Models' (PCM) and 'Rating Scale Models' (RSM)," 14(3) *Rasch Measurement Transactions* 768.

Linacre, J. M. (2007) "Rasch and Continuous Variables," 21(1) *Rasch Measurement Transactions* 1088.

Linacre, J. M. (2013) "Reliability, Separation and Strata: Percentage of Sample in Each Level," 26(4) *Rasch Measurement Transactions* 1399.

Linacre J. M. (2015) *Rasch Measurement Forum.* Available at http://raschforum.boards.net/thread/372/continuous-items

Linacre, J. M. (2016). *Winsteps*® [computer program]. Beaverton, Oregon: Winsteps.com

Linacre, J. M. & W. P. Fisher (2012) "Harvey Goldstein's Objections to Rasch Measurement: A Response from Linacre and Fisher," 26(3) *Rasch Measurement Transactions* 1383.

Mackie, L. (2013) *Law for Life Legal Capability for Everyday Life Evaluation Report*. London: The Gilfillan Partnership.

Marais, I. & D. Andrich (2008) "Effects of Varying Magnitude and Patterns of Local Dependence in the Unidimensional Rasch Model," 9(2) *Journal of Applied Measurement* 1.

Marais, I. (2013) "Local Dependence," in K.B. Christensen, S. Kreiner & M. Mesbah, eds., *Rasch Models in Health.* Hoboken, NJ: Wiley.

Martinez-Martin, P. & M. J. Forjaz (2012) "How to Evaluate Validation Data", in C. Sampaio, C.G. Goetz & A. Schrag, eds., *Rating Scales in Parkinson's Disease: Clinical Practice and Research.* Oxford: Oxford University Press.

Murayama, M. (2007) "Experiences of Problems and Disputing Behaviour in Japan," 14 *Meiji Law Journal* 1.

Nelson, L. R. & M. L. Furst (1972) "An Objective Study of the Effects of Expectation on Competitive Performance," 81 *Journal of Psychology* 69.

Nunnally, J. C. (1978) *Psychometric Theory,* 2d ed. New York: McGraw-Hill.

OECD (Organisation for Economic Co-Operation and Development) (2003) *Learning for Tomorrow's World.* Paris: OECD.

OECD and Open Society Foundations (2016) *Leveraging the SDGs for Inclusive Growth: Delivering Access to Justice for All.* Paris: OECD and Open Society Foundations.

Office for National Statistics (2015) *Harmonised Concepts and Questions for Social Data Sources: Primary Principles – Demographic Information, Household Composition and Relationships (v3.1).* Newport: Office for National Statistics.

Parle, L. J. (2009) *Measuring Young People's Legal Capability*. London: Independent Academic Research Studies and PLEnet.

Persson, C.U., K.S. Sunnerhagen & A. Lundgren-Nilsson (2014) "Rasch Analysis of the Modified Version of the Postural Assessment Scale for Stroke Patients: Postural Stroke Study in Gothenburg (POSTGOT)," 14 *BMC Neurology* 134.

Petrillo, J., S.J. Cano, L.D. McLeod & C.D. Coon (2015) "Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples," 18(1) *Value in Health* 25.

Pleasence, P. & N. J. Balmer (2014) *How People Resolve Legal Problems.* London: Legal Services Board.

Pleasence, P., N. J. Balmer & C. Denvir (2015) *How People Understand and Interact with the Law*. London: Legal Education Foundation.
Pleasence, P. & N. J. Balmer (forthcoming) *Legal Needs Surveys and Access to Justice: A Guidance Document*. Paris: OECD and OSF.
Pleasence, P., C. Coumarelos, S. Forell & H. McDonald (2014) *Reshaping Legal Assistance Services: Building on the Evidence Base.* Sydney: Law and Justice Foundation of New South Wales.
Prieto, L., J. Alonso & R. Lamarca (2003) "Classical Test Theory Versus Rasch Analysis for Quality of Life Questionnaire Reduction," 1 *Health and Quality of Life Outcomes* 27.
RUMM Laboratory (2013a), *Interpreting RUMM2030 – Part III: Estimation and Statistical Techniques.* RUMM Laboratory Pty Ltd.
RUMM Laboratory (2013b) *Interpreting RUMM2030 - Part I: Dichotomous Data.* RUMM Laboratory Pty Ltd.
Salzberger, T. (2010) "Does the Rasch Model Convert an Ordinal Scale to an Interval Scale?" 24 *Rasch Measurement Transactions* 2.
Sandefur, R. L. (2007) "The Importance of Doing Nothing: Everyday Problems and Responses of Inaction," in Pleasence, P. Buck, A. and Balmer, N.J. (eds.) *Transforming Lives: Law and Social Process.* Norwich, UK: TSO.
Sander, P. and L. Sanders (2003) *Measuring Confidence in Academic Study.* Cardiff: University of Wales Institute.
Schwarzer, R. & R. Fuchs (1995) **"**Self-Efficacy and Health Behaviours," in Conner, M. & Norman, P., eds., *Predicting Health Behaviour: Research and Practice with Social Cognition Models.* Buckingham: Open University Press.
Sen, A. K. (1999), *Development as Freedom*. Oxford: Oxford University Press.
Sen, A. K. (2002), *Rationality and Freedom*. Cambridge, MA.: Belknap Press.
Sen, A. K. (2010), *The idea of Justice*. London: Penguin, London.
Sherr, A., R. Moorhead & A. Paterson (1994) *Lawyers: The Quality Agenda.* London: Legal Aid Board.
Shultz, M. M. & S. Zedeck, S. (2011) "Predicting Lawyer Effectiveness: Broadening the Basis for Law School Admission Decisions," 36(3) *Law & Social Inquiry* 620.
Social Research Center (2012) *Legal Issues: Needs of Population of Legal Services and Practiced Ways of Meeting Those Needs.* Dushanbe: Social Research Center.
Smith, E. V. (2002) "Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Components Analysis of Residuals," 3 *Journal of Applied Measurement* 205.
Smith, R. (2014) *Digital Delivery of Legal Services to People on Low Incomes*. London: Legal Education Foundation.
Smith, W., A. Patel, P. McCrone, H. Jin, B. Osumili & B. Barrett (2016) "Reducing Outcome Measures in Mental Health: A Systematic Review of the Methods", 25(5), *Journal of Mental Health* 461.
Strecher, V. J., B. M. DeVellis, M.H. Becker & I. M. Rosenstock (1986) "The Role of Self-Efficacy in Achieving Health Behavior Change," 13 *Health Education Quarterly* 73.
Streiner, D.L., G.R. Norman & J. Cairney (2015) *Health Measurement Scales. A Practical Guide to Their Development and Use (Fifth Edition).* Oxford: Oxford University Press.
Susskind, R. (2008) *The End of Lawyers? Rethinking the Nature of Legal Services.* Oxford: Oxford University Press.
Tang, Y. & H. Tseng (2013) "Distance Learners Self-Efficacy and Information Literacy Skills," 39 *The Journal of Academic Librarianship* 517.
Tennant, A. & P. G. Conaghan (2007) "The Rasch Measurement Modal in Rheumatology: What is it and Why use it? When Should it be Applied, and What Should One Look for

in a Rasch Paper?" 57 *Arthritis and Rheumatism* 1358.
Teresi, J. A., M. Kleinman K. & Ocepek-Welikson (2000) "Modern Psychometric Methods for Detection of Differential Item Functioning: Application to Cognitive Assessment Measures," 19 *Statistics in Medicine* 1651.
TNS-BMRB (2013) *Community Life Survey: Summary of Web Experiment Findings*. London: TNS-BMRB.
Vealey, R. S. & M. A. Chase (2008) "Self-Confidence in Sport," in Horn, T., ed., *Advances in Sport Psychology.* Champaign, Il: Human Kinetics.
Vincent, J.I., J.C. MacDermid, G.J.W. King & R. Grewal (2015) "Rasch Analysis of the Patient Rated Elbow Evaluation Questionnaire," 13 *Health and Quality of Life Outcomes* 84.
Ware, J., M. Kosinski & S. D. Keller (1996) "A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity," 34(3) *Medical Care* 220.
Weisberg, H. F. (2005) *The Total Survey Error Approach: A Guide to the New Science of Survey Research.* Chicago, Il: University of Chicago Press.
Wolfe, E.W. & E.V. Smith (2007) "Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part II – Validation Activities," 8(2) *Journal of Applied Measurement* 204.
Wright, B. D. (1992) "Raw Scores Are Not Linear Measures: Rasch vs. Classical Test Theory Comparison," 6(1) *Rasch Measurement Transactions* 208.
Wright, B. D., & J. M. Linacre (1989) "Observations are Always Ordinal; Measurements, However, Must be Interval," 70(12) *Archives of Physical Measurement and Rehabilitation* 857.
Wright, B.D. (1998) "Rating Scale Model (RSM) or Partial Credit Model (PCM)?," 12(3) *Rasch Measurement Transactions* 641.
Wright, B. D. & G. N. Masters (2002) "Number of Person or Item Strata," 16(3) *Rasch Measurement Transactions* 888.