What women want: a critical appraisal of approaches to evaluating voluntary sector women's services

Kevin Wong *is Reader in Community Justice and Associate Director, Policy Evaluation and Research Unit (PERU), Manchester Metropolitan University.* ***Rachel Kinsella*** *is a freelance researcher and a former member of the team at PERU.* ***Caroline O'Keeffe*** *is a freelance criminal justice researcher as well as a psychotherapist in private practice and mindfulness teacher.* ***Linda Meadows*** *is a freelance researcher with an interest in criminal justice policy and practice.*

This article is the product of a collective approach, rooted in decades of evaluation and research experience, seeking evidence of impact of services for justice-involved women. In conducting evaluations, we have been constrained, to some extent, by the research commissioners seeking evidence that meets *their* needs, that is validation of public policy and justification for public investment. We believe this to be a product of what has been coined as the doctrine of new public management (NPM) which has dominated UK public service since the 1980s and applies corporate performance management frameworks and resource allocation methods to public services. In the context of voluntary sector specialist women's services, this results in an over-fixation with reducing reoffending as the end goal and the use of randomised control trials as the 'holy-grail' of evaluation. We argue that, freed from the shackles of such approaches, it is possible to realise greater benefits for all — commissioners, service providers and, importantly, justice-involved women — through more nuanced evidence gathering. To this end, we argue for applying a scientific realist approach to evaluating women's services, one which starts with: 'what works, for whom, in what circumstances?'. We show that it is only by acknowledging the complexity and changeability of social programme implementation and delivery — the interplay between delivery mechanisms, context and outcomes — and recognising the value of co-production and peer research that we can hope to arrive at an approach to evaluation that actually assists in service improvement and adaptation.

Our colleague, the late Professor Paul Senior provided the template for collectively authoring this paper, which started with the dull but functional title of 'Challenges of evaluating women's services'. Before retirement, as Co-Editor of the British Journal of Community Justice, Paul gathered with colleagues in a Westmoreland hotel. Over two days they engaged in a dialogue and produced 'Imagining Probation in 2020: hopes, fears and insights'.[1] Paul always had a penchant for the grand, but then he was generating content for an entire issue. Our purpose for this sole article was more modest. We no longer work/are affiliated to Sheffield Hallam University (SHU), but it was SHU and Paul that brought us together. So, we returned to the SHU 'canteen' for a day and relaxed, trusting that our contributions would be considered and given due attention regardless of how outlandish. Drawing on our collective several decades worth of evaluation and research experience, we were constructively critical,[2] acknowledging that research is reflexive: at the researcher level; through the politics embedded in the research; and 'the social conditions and techniques of production of the scientific object' (p.441).[3] We took verbatim notes and recorded our discussion. Themes, sub-themes and patterns emerged as we read and re-read the data and listened and re-listened to the recording,[4] guided by the research questions: *What are the challenges of evaluating voluntary sector women's services?* And *how can these be addressed?* Themes and sub-themes were refined for coherence. This paper therefore presents our responses to these questions

1. Senior, P., Ward, D., Burke, L., Knight, C., Teague, M., Chapman, T., Dominey, J., Philips, J., Worrall, A., Goodman, A. (2016). The essence of probation. *British Journal of Community Justice,* 4, 9-27.
2. McWilliams, W. (1980). Management Models and the Bases of Management Structure, South Yorkshire Probation Service Research Unit: Discussion Paper, Series No 26.
3. Wacquant, L. (2011). From 'Public Criminology' To the Reflexive Sociology of Criminological Production and Consumption: A Review of Public Criminology? by Ian Loader and Richard Sparks (Abingdon, Oxon: Routledge, 2010). *The British Journal of Criminology,* 51(2), 438-448.
4. Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology, *Qualitative Research in Psychology,* 3(2), 77 - 101.

contextualised by literature from evaluation, criminology and public service administration and management.

Our paper proceeds thus. We start by defining evaluation and identify factors which hinder the transfer of learning from the evaluation of voluntary sector women's services (hereafter referred to as women's services) into policy and practice. We then examine the appropriateness of applying reducing reoffending as the outcome measure for such provision, proposing an alternative to determining 'impact'. We follow this by proposing that scientific realism offers a more appropriate evaluation approach to facilitating women's service improvement and adaptation. We reflect on our positionality as researchers, our prior experiences, commitment and principles in evaluating women's services and the benefits and challenges of justice-involved women as peer researchers. We conclude with recommendations for policy-makers and commissioners.

What is evaluation good for?

The impact of evaluation is elusive. Ultimately, it's purpose is to determine the value of a treatment or programme, 'to improve or refine the evaluand (formative evaluation) or to assess its impact (summative evaluation)' (p546).[5] Many of our evaluations have been commissioned by government: the Ministry of Justice, HM Prison and Probation Service, the Youth Justice Board, the Home Office; as well as local government, probation and prisons, where the purpose of policy evaluation is to 'systematically investigate the effectiveness of policy interventions, implementation and processes, and to determine their merit, worth, or value in terms of improving the social and economic conditions of different stakeholders.'[6] It's worth noting that our commitment to evaluation (rather than research) rests on its applied nature, a belief that we can draw an intellectual line between our findings, recommendations and policy/programme refinement. Recommendations from our evaluations of services for justice-involved women have sought service commissioning attuned to how women actually engage with services, rather than how commissioners would like them to; policies sympathetic to this; and service adaptations leaning into what women want more and less of.[7] Reflecting on the tenuousness of the described intellectual line, we note the aphorism, attributed to Einstein (but likely apocryphal) that 'insanity is doing the same thing over and over and expecting different results'. Our reporting (ibid) spans several years where broadly the same findings and recommendations have emerged, while at the same time we have observed limited if any change. While frustrating, we have not gone insane. We take solace from the assertion that evaluation offers enlightenment, that research influences policy through ideas rather than data, research is unlikely to produce facts that change policy-making,[8] instead research works through 'knowledge creep',[9] 'through the drip, drip, drip of enlightenment'.[10] In other words, it's a slow process of absorption. And yet, while acknowledging this snail-like pace, we still find it hard to reconcile that the accumulation of knowledge about justice-involved women which we and many other researchers have contributed to — which for example underpinned the government's Female Offender Strategy,[11] has yet to fully find its way into commissioning and practice. Certainly, the spectre of chronic and long-term underinvestment in public services (including women's services) which we have consistently found in our own evaluations stands out as an inhibitor for knowledge application. Without additional resources, effecting change is a struggle. The proposition that it may be possible to do more for less rings hollow after nearly a decade and a half of financial austerity stemming from

We take solace from the assertion that evaluation offers enlightenment, that research influences policy through ideas rather than data.

5. Lincoln, Y., and Guba, E. (1986). Research, Evaluation and Policy Analysis: Heuristics and Disciplined Enquiry, *Policy Studies Review* 5(3), 546 - 565.
6. Government Social Research Unit (2007). The Magenta Book: guidance notes for policy evaluation and analysis. HM Treasury.
7. Kinsella, R., Meadows, L., O' Keeffe, C., Wong, K. (2023). Evaluation of the 'wrap around service' for the mayor's office for policing and crime (MOPAC), Manchester Metropolitan University. Unpublished. Kinsella, R., Clarke, B., Lowthian, J., Ellison, M., Kiss, Z., Wong, K. (2018). Whole System Approach to Women Offenders Final Evaluation Report, Manchester. Manchester Metropolitan University. O' Keeffe, C., Ellingworth, D., Lowthian, J., Clarke, B., Wong, K. (2016) Evaluation of the Whole System Approach for Women Offenders Progress Report. Sheffield Hallam University.
8. Weiss, C., Bucuvalas, M. (1980). Social Science Research and Decision-Making. Newbury Park: Sage.
9. Weiss, C., (1987). The circuitry of enlightenment. *Knowledge Creation Diffusion Utilisation,* 8, 274 - 281.
10. Pawson, R., Greenhalgh, T., Harvey, G., and Walshe, K. (2005) Realist review—a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research and Policy,* 1, 21-34.
11. Ministry of Justice (2018). *Female Offender Strategy.* London: Ministry of Justice.

the 2008 banking crisis.[12] Less well recognised amongst policy-makers and commissioners' thinking is that women's services (and their commissioning) are 'complex systems thrust amidst other complex systems'.[13] The whole system approach to women's services, endorsed in England and Wales as *the model* of women's services is not a closed system, instead it is buffeted, hampered and crowded out by other complex systems.[14] Change can easily be thwarted by complexity. For our part, the neat linearity of causal change: we produce findings/recommendations; policy-makers and practitioners receive them as precision tools; they use them wholly or in part, to tune and fix what's wrong, become blunted and rendered unusable by complexity. We are (naturally) sympathetic to the notion that:

'The relationship between evidence and policy is far from straightforward. Perspectives range from the idealism of 'evidence-driven policy making' (where evidence sets the agenda and drives policy choices) to the pessimism of 'policy-based evidence' (where evidence is sought simply to legitimize pre-set policies).'[15]

By and large, our encounters with the commissioners of women's services and evaluations have persuaded us towards an optimistic iteration of this relationship. They have been as keen as us for evidence to guide what they do. The critical question is what type of evidence can best facilitate this.

What are women's services for?

Evaluation aims, objectives and questions listed in invitations to tender, (necessarily) focus evaluative effort but also box-in evaluators. Over recent years, we have come to the view that they give rise to dashed hopes of 'clinching evidence',[16] invest in X will return reduced recidivism of Y percent. The charge by some public administration and management scholars that evaluation is part of the bundle of practices constituting the doctrine of new public management NPM which has dominated UK public services since the 1980s has resonance.[17] Other NPM practices that we've observed and contributed to (over the last 30 years) which we view as the triumph of management consultancy on public services include: private sector management practices; explicit standards and performance measures; output controls; disaggregation of public services into corporatised units; competition and marketisation; and discipline and parsimony in resource use.[18] After years of faithfully searching for the holy grail of clinching evidence and explaining (apologising) why our efforts have failed, we admit to a disenchantment with a justice policy orthodoxy that is NPM writ large. Albeit recognising that NPM was itself a response to the perceived failings of old public management.[19] For justice-involved women, this manifests as firstly an over-fixation with reducing reoffending as the end goal of services and secondly, an unthinking rush towards randomised experimentation as the only evaluation method that can confirm their worth.

Without additional resources, effecting change is a struggle.

We examine the first here and the second in the next section. One of the authors has written for HM Inspectorate of Probation (HMIP) on why reducing reoffending is not the only outcome for probation and services that work with people with convictions.[20] Below, we apply the central argument to women's services.

A one-size fits all approach to outcome measurement — based principally on the proven rate of reoffending (while strategically and symbolically important) — is unlikely to be sufficiently fine-grained and nuanced to reflect the complex reality of [women and women's services]. The plurality of providers, the different services/functions [that they perform and the different changes in women] that these services are intended to bring about cannot be adequately captured in a simple binary (reoffended or not reoffended) and frequency (if so, how often) measure.' (p.4)[21]

12. Fox, C., Albertson, K., and Wong, K. (2013). Justice Reinvestment: Can the Criminal Justice System Deliver More for Less? London: Routledge.
13. Pawson, R., Greenhalgh, T., Harvey, G., and Walshe, K. (2005). Realist review—a new method of systematic review designed for complex policy interventions. *Journal of Health Service Research and Policy,* 1, 21 - 34.
14. See footnote 11.
15. Cairney, P. (2019). 'Evidence and policy-making' in Boaz, A., Davies, H., Fraser, A. and Nutley, S. (Eds) What Works Now? Evidence-Informed Policy and Practice, Bristol: Policy Press.
16. Hough, M. (2010). Gold standard or fool's gold? The pursuit of certainty in experimental criminology. *Criminology & Criminal Justice,* 10(1), 11-22.
17. Gruening, G. (2001). Origin and Theoretical Basis of New Public Management. *International Public Management Journal,* 4, 1 - 25.
18. Hood, C. (1991). A Public Management for All Seasons? *Public Administration,* 69(10), 3 - 19.
    Osborne, S. (2006). The New Public Governance? 1, Public *Management Review,* 8(3), 377 - 387.
19. Hood, C. (1991). A Public Management for All Seasons, *Public Administration,* 69(1), 3 - 19.
20. Wong, K. (2019). If reoffending is not the only outcome, what are the alternatives? HM Inspectorate of Probation Academic Insights 2019/07.
21. Adapted from Wong, K. (2019:4) see footnote 20.

This begs the question 'What are women's services for?'

Our close examination suggests they take as their starting point the needs of the woman — 'what has happened to you and what do you need' rather than 'what do you need to desist from offending?' In ethos then women's services enable women to become proactive in identifying their own priorities for change. With the relationship between case worker and woman being key to this, that is the craft of working with justice-involved people.[22] And what are those needs?[23] Summary of the literature is still pertinent and resonates with our own more recent work.[24] This includes: '…unmet needs in relation to education, training and employment, health (including mental health), housing and income'; sexual and violent victimisation; high rates of substance misuse, especially opiates, amongst female offenders; poverty and financial difficulties; with women's financial situations '…further strained by their having sole responsibility for dependent children.'

The key argument is that alternative outcomes (to reducing reoffending) should be ones that the women value and which enable them to make the micro-changes necessary to progress their lives. These outcomes may proffer limited gain for justice policy but cumulatively garner significance for health improvement, social capital and other public policy goals, eventuating a reduced reliance on state provision. These outcomes should take primacy. For an exposition of what these might be, see the interview with Lisa Dando and colleagues in this publication. These arguments for alternative outcomes — applied here to women's services — are part of a broader movement attempting to grapple with the complexity of service delivery within complex systems.[25] While others are pioneering relational approaches to public service delivery — 'the liberated method' proffering effectiveness — serving the needs of service users rather than efficiency,[26] the legacy of NPM. Useful learning from these initiatives can be applied to women's services.

It should be noted of course that women's services struggle for funding. Their reliance on government sources leaves them treading a difficult path, to avoid being complicit in enforcement, where missed appointments trigger breach actions by probation.

What works for whom in what circumstances…?

The complexity of women's services and policy/delivery landscape they inhabit steers us to advocate for a scientific realist approach to evaluating women's services. In realist evaluation the question of 'what works' with its seductive simplicity becomes the more nuanced 'what works, for whom, in what circumstances, to what extent'. As far as we know, few if any women's service evaluations have explicitly adopted a realist approach — our own included. Realist evaluation recognises the complexity and changeability of social programme implementation and delivery, that the *mechanisms* — which underpin women's services, (practitioners interactions with women, women's reasoning, the processes which affect their behaviours and so on) are affected by *context* (women's characteristics, local infrastructure, socio-economic conditions, access to services, the requirements of other services, the co-operation or lack of co-operation of services; family and peer relationships). They generate *outcomes* intended and unintended: Women secure housing close to supportive family, have better access to their children; however, it takes three rather than one bus to attend their appointments with probation and they begin to miss them. It assumes that services will be optimal for some women but not others, but this could alter if circumstances change. Realist evaluation works with the untidy non-linear complex messiness of the social world as it is, rather than the tidy linear version in policy makers and commissioners'

For justice-involved women, this manifests as firstly an over-fixation with reducing reoffending as the end goal of services.

22. Hough, M. (2010). Gold standard or fool's gold? The pursuit of certainty in experimental criminology. Criminology & Criminal Justice, 10(1).
23. Gelsthorpe, L., Sharpe, G., Roberts, J. (2007). Provision for Women Offenders in the Community. Fawcett Society.
24. Kinsella, R. Clarke, B. Lowthian, J. Ellison, M. Kiss, Z. Wong K. (2018). Whole System Approach to Women Offenders Final Evaluation Report, Manchester. Manchester Metropolitan University.
25. French, M., Hasselgreave, H., Wilson, R., Lowe, T., & Hawkins, M. (2023). Harnessing Complexity for Better Outcomes in Public and Non-profit Services. Policy Press.
26. Smith, M. (2023). The Liberated Method - Rethinking Public Service. Changing Futures. Northumbria.

heads. If we commission service A to do B + C then Y will happen. There are no magic bullets to deal with the complexity of marginalised women's lives. A service working effectively with a subgroup of women in one area may fare poorly with a similar subgroup elsewhere but instead achieve success with a different subgroup. Women's services need to adapt to different conditions, a role that realist evaluation is designed to support. Realist evaluation describes a 'realist' evidence-based pathway chain which through theory elicitation, then testing facilitates programme targeting and programme improvement.[27] This is a heuristic; however, it is instructive, the point of evaluation here is programme improvement, enabling service(s) to adjust and refine provision. Borrowing from Pawson and colleagues, women's services are:

> *'…dynamic complex systems thrust amidst complex systems, relentlessly subject to negotiation, resistance, adaptation, leak and borrow, bloom and fade.' (p23)*[28]

Even if women's services themselves don't change, things around them do: a new funding regime; new national and local policies/strategies emerge; women's needs change; agencies that women are referred to cease operating. Adaptation is constant, at times more urgent at other times less all embracing.

Women's services struggle for funding. Their reliance on government sources leaves them treading a difficult path.

We return to our earlier point about policy orthodoxy. Randomised control trials (RCTs) have become the gold standard of evidence-based policy (EBP) in England and Wales and firmly embedded in the What works movement and What works Centres established since 2001.[29] Our several decades experience accords with White's four waves of the evidence revolution: the NPM results agenda; the rise of impact evaluations specifically RCTs since 2000s; systematic reviews of RCT evidence; and institutionalising evidence use through knowledge brokers, the What Works Centres.[30] In 2010, the march towards randomised experimentation for justice programmes generated lively debate within academic criminology.[31] The arguments for and against experimentation still apply, however, within government, the door seems firmly shut, the RCT horse has bolted and is on the loose. Let us be clear, we are not against RCTs, however, we have reservations about its widespread and indiscriminate application, such as for women's services.

Pawson's description of the orthodox evidence based policy (EBP) pathway is instructive, it starts with policy instigation, moves to programme management, onto demonstration project and then to full-scale RCT over a two-five year time frame.[32] Admittedly, this again is simplified for illustrative purposes. The rush to evidence by what works centres where a breathless two years has become the norm (for example see Youth Endowment Fund commissioned evaluations) and where the comparatively pedestrian five years is eschewed is a tad perplexing.[33] Perplexing because the timeframe for evidence-based medicine (EBM), upon which the foundations of EBP have been built, is ten-fourteen years for drug development where a full scale Phase III RCT is at the end of a long chain of research activities.

The uncritical importing of EBM to EBP without paying sufficient attention to purpose and context is misguided.[34] Pawson's exposition of the differences is instructive.

> *'All the design features of drug RCTs are interrogated and fixed prior to testing. The net effects of drug RCTs speak to ideal*

27. Pawson, R. (2017) Evidence-based Medicine & Evidence-based Policy: The world's most perfectly developed method & the 79-pound weakling? University of Leeds.
28. Pawson R, Greenhalgh T, Harvey G, Walshe K. (2005) Realist review—a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy,* 10(1), 21-34.
29. Haynes, L., Service, O., Goldacre, B., and Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. Cabinet Office.
30. White, H. (2019). 'The twenty-first century experimenting society: the four waves of the evidence revolution'. *Palgrave Communication* 5(47).
31. Sherman, L. (2009). Evidence and liberty: the promise of experimental criminology, *Criminology and Criminal Justice,* 9(1), 5 - 28. Tilley, N. (2009). Sherman vs Sherman: realism vs rhetoric, *Criminology and Criminal Justice,* 9(2),135-134.
32. Pawson, R. (2006). Evidence-based policy: A realistic perspective. Thousand Oaks, CA: Sage.
33. These can be found here: https://youthendowmentfund.org.uk/funding/evaluations/
34. Pawson, R. (2017) Evidence-based Medicine & Evidence-based Policy: The world's most perfectly developed method & the 79-pound weakling? University of Leeds.

conditions which are reproducible. All the design features of social programme RCTs are improvised. The net effects of social programme RCTs are ad hoc partial artefacts.'

This difference can perhaps be explained by the philosophical histories and underpinnings of natural sciences and social sciences observed by Rosenberg.[35]

'The natural sciences have a much larger body of well-established, successful answers to questions and well-established methods for answering them…many of the basic philosophical questions about the limits and the methods of the natural sciences have been set aside in favor of more immediate questions . . . . The social and behavioral sciences have not been so fortunate…there is no consensus on the questions that each of them is to address, or on the methods to be employed.' [36]

Except of course, policy makers and government have pressed on, determining the questions and methods that evaluators have to work within. This has always been so. Whoever pays, calls the evaluation tune. As evaluators we have striven to deliver what commissioners have asked for within reason and without comprising principles of being guided by the evidence and avoiding harm. This paper is a rare opportunity to step back from our day-to-day role and reflect on what we have learned from our striving.

Which brings us back to the question why has EBP adopted EBM wholesale? We view this is an unadmitted inferiority amongst social policy makers and researchers (ourselves included) coupled with a longing for rigour and perceived certainty, the clinching evidence that EBM appears to provide. Like a seventies teenager admiring and then appropriating their older sibling's achingly cool LP collection. And yet, not so perhaps. Pawson's coda to the EBM pathway for drug development looks at what happens post regulatory approval (after a confirmatory Phase III trial).[37] When drugs are used in the open real-world system compared to the closed world of an RCT: compliance with treatment will worsen; the limited co-morbidities of patients in RCTs will differ from the general population; there will be greater heterogeneity in outcomes; unintended consequences will emerge; drug resistance will occur. In reality, EPM also has to wrestle with the uncertainty, complexity and messiness of the real world. At this point the EPM pathway with a defined end, becomes an evaluation cycle where the Phase III trials 'should be understood not so much as 'final arbiters' but as 'way stations' representing current distillations of knowledge.' (p16)[38]

Women's services need to adapt to different conditions, a role that realist evaluation is designed to support.

We come full circle. Above we have laid out the challenges of evaluating women's service. Here, we proffer solutions, recommendations for policy makers and commissioners — a reflex, conditioned by decades of evaluation.

We recommend a return to first principles, specifically those provided by two essential government texts: Government Social Research (GSR) ethics guidance,[39] and the Magenta Book, Central Government Guidance on evaluation.[40] They are summarised below.

Outcomes for women's services. The outcomes and measurement of such through evaluation should align squarely with what women want and what women's services do rather than what commissioners would like them to do. This necessarily requires a co-production approach — which is clearly supported by *Principle 1 Research should have a clear user need and benefit* and *Principle 5 Research should enable participation of the groups it seeks to represent.*[41]

35. Rosenberg, A. (2012) Philosophy of Social Science, Boulder: Westview.
36. See footnote 26: Rosenberg, A. (2021).
37. Pawson, R. (2017). Evidence-based Medicine & Evidence-based Policy: The world's most perfectly developed method & the 79-pound weakling? University of Leeds.
38. Pawson, R. (2017). Evidence-based Medicine & Evidence-based Policy: The world's most perfectly developed method & the 79-pound weakling? University of Leeds.
39. Government Social Research. (2021). GSR Professional Guidance: Ethical assurance for Social and Behavioural Research in Government. GSR.
40. HM Treasury. (2020). *Magenta Book Central Government Guidance on Evaluation.* https://assets.publishing.service.gov.uk/media/5e96cab9d3bf7f412b2264b1/HMT_Magenta_Book.pd
41. Government Social Research. (2021). GSR Professional Guidance: Ethical assurance for Social and Behavioural Research in Government. GSR.

Research method. Research paradigm wars. Less a war — given the current hegemony of experimental design — and more skirmish. Realist evaluation approaches have the potential to provide a more sympathetic way of doing what evaluation can do well, not provide clinching evidence but instead facilitate service improvement and adaptation and provide a better understanding about what changes/outcomes have occurred and why. Figure 2.4 of the Magenta Book (p33) sets out a decision tree to select the most appropriate method for impact evaluation based on intervention type.[42] The characteristics of women's services falls squarely within the conditions for adopting a theory-based evaluation approach, which realist evaluation fulfils. Moreover, a more careful consideration of research method operationalises *Principle 5 Research should enable participation of the groups it seeks to represent.* The guidance confirms this 'not only helps to ensure the effective dissemination and impact of research findings, but also is an important step in determining the most appropriate and effective research methods.' (p.5) [43]

Peer research. As one manifestation of co-production in research and operationalising the potential turn in public service administration towards the collaboration ethos of new public governance,[44] peer research involving justice-involved women has much to commend it. As a method it clearly upholds ethics *Principle 5 Research should enable participation of the groups it seeks to represent.*[45] Since two of the authors trialled it over twenty years ago as a then novel approach, it has since become more widely adopted. However, effecting change, in this case evaluation practice, requires investment of additional resources, but also care in implementation and a willingness to forgo control as a professional researcher and share this.[46]

How to do this are provided in other texts precluded from inclusion here by word limit.[47] Our experiences of these co-produced practices, specifically, training and supporting peer researchers to undertake research with women has demonstrable benefits: reducing the gap between researcher and researched; women can be themselves, say what is true for them rather than a filtered version; offering a richer, alternative insight that a 'professional researcher' may not be able to elicit. For the peer researchers, it allowed them to forge a professional pathway from their experiences, it gave them purpose, self-worth, status and an alternative identity. That contemporary evaluation commissioners may now favour this indicates a shift from the commissioning and evaluation landscape of twenty years ago. Of course, being a peer researcher is not an unalloyed good. The challenges they experience — confusion of identity; going from the 'high' and self-affirming experience of being a researcher to returning to prison and resuming their status as prisoner speaks to the challenges of such peer roles.[48] Careful attention to managing these contradictions is important.

We end by invoking complementary ideas from two twentieth century social scientist pioneers (Robert Merton and Donald Campbell) to support a call for a vigorous debate about how best to evaluate women's services, ultimately, the purpose of this paper. Science (in this case evaluation) is advanced by organised scepticism and does not depend on elite consensus and infallible evidence.[49] And objectivity in science, gathers through social processes, where scientists compete, check and challenge each other's interpretations.[50] We invite policy makers, practitioners and other evaluators to join with us in such discourse.

42. HM Treasury. (2020). Magenta Book Central Government Guidance on Evaluation. https://assets.publishing.service.gov.uk/media/5e96cab9d3bf7f412b2264b1/HMT_Magenta_Book.pd
43. Government Social Research (2021). GSR Professional Guidance: Ethical assurance for Social and Behavioural Research in Government. GSR.
44. Osborne, S. (2006.) The New Public Governance? 1. *Public Management Review* 8 (3) pp.377–387.
45. Government Social Research (2021). GSR Professional Guidance: Ethical assurance for Social and Behavioural Research in Government. GSR.
46. O'Keeffe, C. (2004). Object and Subject: The Challenges of Peer Research in Community Justice in British *Journal of Community Justice* Vol.3(1). Buck, G., Harriott, G., Ryan, K., Ryan, N., and Tomczak, P. (2020). All our justice: People with convictions and 'participatory' criminal justice, in The Routledge Handbook of Service User Involvement in Human Services Research and Education Edited by Hugh McLaughlin, Peter Beresford, Colin Cameron, Helen Casey, Joe Duffy.
47. O'Keeffe, C. (2003). Moving Mountains: Identifying and Addressing Barriers to Employment, Training and Education from the voices of women (ex) offenders. Sheffield: SHU Press.
48. O'Keeffe, C. (2003) Moving Mountains: Identifying and Addressing Barriers to Employment, Training and Education from the voices of women (ex) offenders. Sheffield: SHU Press.
49. Merton, R. (1973). The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press.
50. Campbell, D., and Russo, J. (1998) Social Experimentation. Thousand Oaks, Sage.