# JOINT TRIALS AND PREJUDICE: A REVIEW AND CRITIQUE OF THE REPORT TO THE ROYAL COMMISSION INTO INSTITUTIONAL CHILD SEX ABUSE

PETER M ROBINSON*

*One of the lesser known tasks of the Royal Commission into Institutional Responses to Child Sexual Abuse was to investigate responses within the legal system to allegations of such abuse, including the procedural and evidentiary rules surrounding joinder of complaints by multiple complainants against the same defendant. The Commission itself commissioned an empirical study and a report on the effects of joinder of charges on jury reasoning and decision making, which, at over 370 pages in length is quite demanding to digest, and, we would argue, open to criticism on methodological and interpretive grounds. This article reviews and critiques the report's methodology and findings, and argues for interpretations and conclusions contrary to those contained in the report, to the effect that the study did provide significant evidence supporting the prejudicial effect of joinder and failed to adequately controvert theories of prejudice through character bias, accumulation prejudice and inter-case conflation of evidence.*

Years of progressive revelation of institutional child sex abuse, and of the manifold ways in which institutions have ignored or concealed it, inevitably culminated in a governmental response. In January 2013, a Royal Commission was appointed with wide-ranging terms of reference empowering it, inter alia, to investigate and recommend appropriate responses by institutions, government and statutory authorities to allegations of sexual abuse of children.[1] The possible governmental responses being investigated included matters of legislative policy relating to the admission of prior conduct and convictions, an area with potential impacts beyond the ambit of the Commission's terms of reference.

As part of its enquiry, the Commission itself commissioned research on the criminal justice system, directed to the process of investigation and prosecution of complaints of child sexual abuse, and these reports have been relied on by the Commission in formulating recommendations in its Criminal Justice Report released on 14 August 2017 ('the CJR'), which then became part of their Final

* School of Business and Law, Central Queensland University, Brisbane, Australia. Email: p.robinson1@cqu.edu.au. This research was conducted with the support of Central Queensland University and Macquarie University.

1 Royal Commission into Institutional Responses to Child Sex Abuse, *Terms of Reference* (13 November 2014) <https://www.childabuseroyalcommission.gov.au/terms-reference>.

Report.[2] The laws for admission of past conduct as evidence of guilt were the subject of an advice on the Australian law by Game, Roy and Huxley and a comparative study by Hamer addressing analogous laws in selected foreign jurisdictions.[3] The same laws underpin the joinder of charges in a single trial, since cross-admissibility is usually the foundation for such joinder,[4] although it may not be a strict legislative requirement.[5]

Joinder is an issue of particular sensitivity in sexual assault trials, especially those involving child victims, since the alleged conduct often occurs in private with no witnesses present, and the alleged offence is arguably characterised by a pattern of repetitive conduct. Jury reasoning in joint and separate trials of child sexual abuse was the subject of an empirical study ('the JT Study') by Jane Goodman-Delahunty, Annie Cossins and Natalie Martschuk ('the JT researchers') whose findings were published in a report ('the Report')[6] for the Commission.

The JT Study makes valuable contributions to the body of knowledge in this field, but since the Report is 376 pages in length, much of it enshrouded in statistical detail, it is relatively inaccessible to all but the most diligent and statistically savvy of lawyers. The present article has the dual goals of reviewing the main findings of the JT Study in relation to joinder and the admissibility of prior conduct in order to make them more transparent to policy makers and researchers, and to critique the methodology and conclusions. In the process of review, several criticisms and alternative interpretations came to light which run counter to the final conclusions of the Report. These are canvassed in detail. The JT Study was quite wide-ranging with many secondary findings, so for present purposes, some simplifications and omissions have been made to focus on aspects directly relevant to the effect of joinder of trials and the use of prior conduct evidence.

2 Commonwealth, Royal Commission into Institutional Responses to Child Sexual Abuse*, Criminal Justice Report* (2017); Commonwealth, Royal Commission into Institutional Responses to Child Sex Abuse, *Final Report: Recommendations* (2017).

3 Tim Game, Julia Roy and Georgia Huxley, *Tendency, Coincidence and Joint Trials* (14 September 2015) Royal Commission into Institutional Responses to Child Sexual Abuse <https://www.childabuseroyalcommission.gov.au/documents/advice-on-tendency-and-coincidence-evidence-and-jo.pdf>; David Hamer, *The Admissibility and Use of Tendency, Coincidence and Relationship Evidence in Child Sexual Assault Prosecutions in a Selection of Foreign Jurisdictions* (March 2016) Royal Commission into Institutional Responses to Child Sexual Abuse <https://www.childabuseroyalcommission.gov.au/getattachment/be2d90fa-f901-43c8-a056-a7005fa0d84c/The-admissibility-and-use-of-tendency,-coincidence>.

4 See *Sutton v The Queen* (1984) 152 CLR 528, 531; *De Jesus v The Queen* (1986) 22 A Crim R 375, 377–8; *Phillips v The Queen* (2006) 225 CLR 303, 307; *Billings v The Queen* [2012] NSWCCA 33 (16 March 2012) [15]–[16]; *Rapson v The Queen* (2014) 45 VR 103, 104; *Young v The Queen* [2015] VSCA 265 (22 September 2015) [4], [6].

5 See Criminal Procedure Act 1986 (NSW) s 29; Criminal Procedure Act 2009 (Vic) ss 193–4; Criminal Code Act 1899 (Qld) ss 567, 568; Criminal Law Consolidation Act 1935 (SA) s 278; Criminal Procedure Act 2004 (WA) ss 133, 134.

6 Jane Goodman-Delahunty, Annie Cossins and Natalie Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study* (May 2016) Royal Commission into Institutional Responses to Child Sexual Abuse <https://www.childabuseroyalcommission.gov.au/getattachment/b268080d-599a-4d44-a9c5-c3f8181bae96/Jury-reasoning-in-joint-trials-of-institutional-ch>.

## I LEGAL BACKGROUND

The potential prejudicial effect of prior conduct evidence has long been recognised by the courts, and led to quite stringent common law tests for admissibility. At one time it was thought that an argument based on propensity was entirely impermissible, so prior conduct evidence could only be admitted for some other purpose.[7] In the United Kingdom, there was once an effort to categorise those permissible purposes, but this approach was ultimately rejected in *Director of Public Prosecutions v Boardman*.[8]

English courts seemed initially to prefer reasoning based on coincidence rather than propensity. In Australia, while inference from a general propensity to commit crimes or a particular type of crime was frowned on, courts were amenable to reasoning based on more specific propensities so long as their probative value outweighed any prejudicial effect.[9] This approach was later adopted in England in *Director of Public Prosecutions v P*.[10] However, continued judicial caution in Australia led to the development of a demanding requirement that the propensity evidence must possess such probative value that, if accepted, it bore no reasonable explanation other than the inculpation of the accused.[11] This test, known as the *Pfennig* test, has been much criticised, although it could be seen as obviating the need for an awkward weighing of probative value against prejudicial effect.[12]

Impetus for reform, spurred by Law Reform Commissions, led to the introduction of Uniform Evidence Acts in the 1990s, which now operate in the Commonwealth, three states and two territories.[13] They set the threshold for prima facie admissibility of tendency and coincidence evidence at 'significant probative value' (ss 97–8), a standard somewhat higher than mere relevance but less than 'substantial',[14] balanced by a further requirement that the probative value must substantially outweigh the prejudicial effect (s 101). The interpretation of 'significant probative value' has led to some tension between interpretations by the Courts of Appeal in New South Wales and Victoria, which may have been

7 *Makin v A-G (NSW)* [1894] AC 57, 65 (Lord Herschell LC); *DPP v Boardman* [1975] AC 421, 438–41 (Lord Morris); *Markby v The Queen* (1978) 140 CLR 108, 116 (Gibbs ACJ).

8 *DPP v Boardman* [1975] AC 421, 439 (Lord Morris), approved by the High Court in *Markby v The Queen* (1978) 140 CLR 108, 116 (Gibbs ACJ).

9 See *Hoch v The Queen* (1988) 165 CLR 292, 301–2 (Brennan and Dawson JJ); *Harriman v The Queen* (1989) 167 CLR 590, 597 (Dawson J), 607 (Toohey J), 613 (Gaudron J).

10 [1991] 2 AC 447, 460–1 (Lord Mackay LC).

11 *Pfennig v The Queen* (1995) 182 CLR 461, 481–2 (Mason CJ, Deane and Dawson JJ) ('*Pfennig*'), applying *Hoch v The Queen* (1988) 165 CLR 292, 294–5.

12 See David Hamer, 'The Structure and Strength of the Propensity Inference: Singularity, Linkage and the Other Evidence' (2003) 29 *Monash University Law Review* 137; Annie Cossins, 'The Legacy of the *Makin* Case 120 Years On: Legal Fictions, Circular Reasoning and Some Solutions' (2013) 35 *Sydney Law Review* 731, 750, and references there cited.

13 *Evidence Act 1995* (Cth); *Evidence Act 1995* (NSW); *Evidence Act 2008* (Vic); *Evidence Act 2001* (Tas); *Evidence Act 2011* (ACT); *Evidence (National Uniform Legislation) Act 2011* (NT).

14 *Lockyer* (1996) 89 A Crim R 457, 459; *AW v The Queen* [2009] NSWCCA 1 (30 January 2009) [47]; *BJS v The Queen* (2013) 231 A Crim R 537, 548 [47].

alleviated by the recent decision of the High Court in *Hughes v The Queen* in favour of not requiring a high degree of similarity.[15]

Western Australia also uses 'significant probative value' as the threshold for prima facie admission, balanced by an exclusionary rule expressed in terms of the public interest.[16] In Queensland, the common law test in *Pfennig* still applies.[17] The South Australian legislation applies the common law approach, without the *Pfennig* proscription, by excluding evidence of discreditable conduct to show a general propensity, but otherwise admitting it if the probative value substantially outweighs its prejudicial effect.[18] In the United Kingdom, there has also been legislative change. The *Criminal Justice Act 2003* of England and Wales allows the admission of evidence of bad character in a number of circumstances, the most general of which are when it is 'relevant to an important matter in issue between the defendant and the prosecution',[19] or when it is 'important explanatory evidence'.[20] There are general discretions to exclude if it would have an adverse effect on the fairness of the trial.[21] Setting the threshold for prima facie admission at mere relevance, albeit with a requirement of importance, has generally loosened the shackles on bad character evidence compared to the common law, despite the exclusionary rule.[22]

Historically, the decision to allow joinder of charges arising from conduct on separate occasions, often involving multiple complainants, turned on cross-admissibility of the evidence across the charges, and that was only likely if the conduct on discrete occasions was admissible as tendency evidence. As *Hughes* shows, even under modern legislation in which cross-admissibility is not a strict requirement for joinder, it continues to exert considerable influence on such decisions.[23]

Within the scheme of the Uniform Evidence Acts, prejudice has an influence at two levels. First, it provides the policy rationale for imposing a higher threshold of admissibility than mere relevance, and secondly, it provides a basis for excluding the evidence despite satisfying that threshold. Despite the legislative advances, there are still urgings in Australia for reform in the direction of leniency towards admission, the most prominent being in the CJR, which recommends that in child sexual abuse cases the threshold for admission be mere relevance to 'an important evidentiary issue', based on the test currently applying in England and Wales.[24] This recommendation is expressly based on a view derived from the Report that

15 *Hughes v The Queen* (2017) 344 ALR 187, 192 [12], 199 [40] ('*Hughes*').

16 *Evidence Act 1906* (WA) s 31A(2).

17 See *R v CBM* [2015] Qd R 165, 175 [40]–[44].

18 *Evidence Act 1929* (SA) s 34P.

19 *Criminal Justice Act 2003* (UK) c 44, s 101(1)(d).

20 Ibid s 101(1)(c).

21 Ibid s 101(3); *Police and Criminal Evidence Act 1984* (UK) c 60, s 78.

22 See Hamer, above n 3, 33.

23 *Hughes* (2017) 344 ALR 187, 191 [6].

24 Royal Commission into Institutional Responses to Child Sexual Abuse, *Criminal Justice Report*, above n 2, Executive Summary, 72; Royal Commission into Institutional Responses to Child Sexual Abuse, *Final Report: Recommendations*, above n 2, 105–6.

'many of the concerns and criticisms [about prejudice] are not well founded'.[25] The soundness of the Report's findings is therefore fundamental to any consideration of the Royal Commission's recommendations.

## II ISSUES ADDRESSED BY THE JT STUDY

In broad terms, the goals of the JT Study were:

1. to determine whether juries in joint trials with multiple complainants and cross-admissible evidence engage in impermissible reasoning due to such joinder; and
2. to examine the effect of judicial directions and question trails designed to eliminate impermissible reasoning.[26]

One of the limitations of previous jury studies in this field was that they failed to distinguish impermissible from permissible reasoning based on evidence of prior conduct, convictions or allegations. Earlier studies leave little doubt that evidence of prior convictions will increase the conviction rate, but whether that is due to rational inference from a tendency pattern or irrational prejudice is unclear.[27] Advocates for the admission of prior conduct argue for its probative value, which would increase the conviction rate for guilty defendants, while opponents would argue that it increases conviction of the innocent, or at least diminishes the threshold of reasonable doubt.[28]

Although judicial cautions about prior conduct evidence have a long history in legal decision-making, precise judicial formulations of the feared prejudice are surprisingly scarce, particularly in Australia. While the JT researchers did consider previous jury research, they relied primarily on a 1976 decision of the United States Court of Appeals for the Fourth Circuit in *United States v Foutz*[29] to identify the forms of potential prejudice. In that case, three sources of possible prejudice were described:[30]

1. Inter-case conflation of evidence;
2. Accumulation prejudice; and
3. Character prejudice.

25 Royal Commission into Institutional Responses to Child Sexual Abuse, *Criminal Justice Report*, above n 2, pt VI, 617.

26 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 71.

27 Ibid 51, 56–7.

28 See, for example, Murphy J in *Perry v The Queen* (1982) 150 CLR 580, 594.

29 United States v Foutz, 540 F 2d 733 (4th Cir, 1976). See Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 45–6.

30 *United States v Foutz*, 540 F 2d 733, 736 (4th Cir, 1976).

Inter-case conflation of evidence was described as the confusion of evidence across charges, while accumulation prejudice involves giving undue weight to the accumulation of evidence across charges.[31] Interestingly, the US Court defined character prejudice in a way that denies the distinction between permissible and impermissible reasoning about propensity used in the JT Study. Its definition of character prejudice (summarised in the Report) as 'the use of evidence from one crime to infer criminality on the part of the defendant [in another crime]' presupposes that evidence of past criminal conduct can never be probative across charges,[32] which does not reflect the Australian position. According to Hamer, this strict approach in United States law has been watered down in practice by multiple qualifications.[33]

For the purposes of the JT Study, the three types of potential prejudice were defined more formally as follows:[34]

> *Inter-case conflation of the evidence*: A type of impermissible reasoning based on substitution of the facts in evidence about one complainant for facts in evidence about another complainant, in a joint trial involving two or more complainants.
>
> *Accumulation prejudice*: A type of impermissible reasoning that accords more weight to evidence than its true value, because multiple charges or multiple witnesses who give evidence against a defendant create the appearance of a stronger case against the defendant than exists in reality.
>
> *Character prejudice*: A type of impermissible reasoning based on the unwarranted inference of criminality in a defendant who is thus considered to deserve punishment because he or she is a bad person.

'Impermissible reasoning' was defined as '[r]easoning that is logically unrelated to the evidence'.[35]

Another important definition is that of the 'joinder effect', which was defined as a 'statistically significant increase in the conviction rate for an offence when it is tried in a joint trial, compared to the conviction rate for the same offence when it is tried in a separate trial'.[36] The article cited in relation to that definition,[37] and the following passage, make it clear that what is being referred to is a comparison of a joint trial with multiple complainants and a separate trial with a single complainant:

> Strictly speaking, the joinder effect describes elevated conviction rates for the focal counts in a joint trial compared to similar counts in separate trials. In trials that involve multiple complainants and a single defendant, the law assumes there

31 Ibid.

32 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 46; ibid.

33 Hamer, above n 3, 72.

34 See Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 13, 17.

35 Ibid 17.

36 Ibid 18.

37 Andrew D Leipold and Hossein A Abbasi, 'The Impact of Joinder and Severance on Federal Criminal Cases: An Empirical Study' (2006) 59 *Vanderbilt Law Review* 349.

> will be some degree of unfair prejudice to the defendant because the jury has heard about the defendant's other criminal misconduct from the evidence of multiple complainants.[38]

However, the Report also frequently refers, confusingly, to a different, undefined joinder effect, which it sometimes calls the 'joinder effect per se'.[39] This involves comparison of a complaint tried in a joint trial to the same complaint tried separately but with the evidence of the other complainants admitted as tendency evidence. This joinder effect, if it occurs, is more difficult to rationally explain because the two types of trial have identical evidence, but it is not the joinder effect which typically creates controversy in practice. In practice, cross-admissibility and joinder go together, so the courts are not called upon to consider severance of the charges without severance of the other complainants' evidence. In the Report, as in practice, the key issue was whether the jury should be allowed to hear the evidence of other offending conduct,[40] which is the issue impacted by the joinder effect as actually defined.

## III METHODOLOGY

The JT researchers eschewed experimental paradigms designed to explore human reasoning in general in favour of studies specifically targeting jury decision-making.[41] They sought to achieve authenticity (or what psychologists would call ecological validity) by simulating real world trial circumstances as much as possible within an experimental setting.[42]

### A *Participants*

For the main study, jurors were recruited from the jury-eligible population through an offer including a $100 incentive for participation.[43] They completed a pre-trial questionnaire when they registered online, to assess their a priori expectations and attitudes.[44] The 1029[45] volunteers (580 women, 449 men) were rather well educated — 47 per cent had tertiary degrees and 13.8 per cent were currently

38 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 251.

39 See, eg, ibid 252–3

40 Ibid 38.

41 Ibid 23, 44, 66.

42 Ibid 268.

43 Ibid 80.

44 See ibid 324–7. Results of this pre-trial questionnaire were used to ensure that observed differences in outcomes were due to experimental effects and not pre-existing biases or attitudes.

45 This is the figure quoted in the narrative of the Report: see Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 22, 24, 39, 74, 93 (Table 3 notes), 244. However, it does not tally with the total of 1031 appearing in Table 2 at 77. For the purposes of Table 1 below, and consequent analyses, Table 2 of the Report was relied on.

undertaking tertiary education — but the spread of ages was balanced (youngest 18 years, oldest 82 years).[46]

## B *Procedure*

They were randomly assigned to attend a video-recorded trial of 45 to 110 minutes duration[47] in which professional actors played the witnesses, complainants and defendants, and real-life judges and barristers played the legal roles. After viewing the trial, they were given 90 minutes to reach a *unanimous* verdict, during which their deliberations were recorded (and later transcribed for analysis).[48] The verdict was then provided by a randomly selected foreman of the jury,[49] and each juror completed a post-trial questionnaire which assessed more specific questions about their conclusions, perceptions of the trial process, cognitive effort and affect (ie emotional reaction to the trial).[50] This questionnaire also asked jurors individually for their personal verdict,[51] a fact which becomes important when the results are analysed because the individual juror verdicts provide a much larger statistical sample than the 'grouped' verdicts of juries. The JT researchers argue that the group verdicts are more valuable because they are more authentic,[52] but the individual verdicts do incorporate effects of the group dynamic because the individual jurors were polled immediately *after* the group deliberations took place.

## C *The Pilot Study*

Prior to the main study, a 300-participant online pilot study was employed to confirm the strength of the evidence in the mock trial scenarios, presented in the pilot study as written scripts.[53] The aim was to have mock trials of complaints with objectively strong, moderate and weak evidence.[54] The strength of the evidence was manipulated by including things like factual discrepancies in the weak case and corroboration in the strong case (for fuller descriptions, see below).[55] The results indicated statistically significant differences in the individual conviction rates for different strengths of evidence, especially with respect to weak versus moderately strong evidence. Weak evidence produced conviction rates on a single count of 24 per cent, while moderately strong evidence produced rates in

46 Ibid 24, 273–4 (Appendix A).

47 The duration of the trial was dependent on the number of witnesses and charges.

48 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 81.

49 Ibid 81.

50 Ibid 79–80, 335–42.

51 Ibid 335.

52 Ibid 44, 66, 270.

53 Ibid 72–3.

54 Ibid.

55 Ibid 84–8.

the range of 50–60 per cent.[56] Strong evidence produced conviction rates in the range of 65 per cent to 75 per cent for a minor indecency offence, but down to 58 per cent for a more serious penetrative offence (compared to 52 per cent for the moderately strong evidence).[57] The JT researchers regarded this as acceptable, on the basis that the major question in a joint trial is the impact of such joinder on a claim with the weakest evidence[58] — a curious argument in view of the fact that the weak case did not become the focal case in the main study.

## D *The Mock Trials*

Scripts for the trials were based on real-life cases.[59] Each allegation was made by a former member of an under-12 boys' soccer team coached by the male defendant, and the charges were laid about 20 years later when the complainants were around 30 years of age.[60]

The trial with moderate-strength evidence was treated as the focal case. This meant that it was the subject of charges in all trials, and effects of joinder and tendency evidence were adjudged by reference to verdicts on those charges.[61] The strong and weak cases acted as tendency evidence in the tendency trials but were also the subject of charges in the joint trials.[62]

The moderate-strength trial involved two counts, one an act of indecency, involving the defendant stroking the complainant's penis and forcing the complainant's hand onto the defendant's penis, and the other an act of unlawful sexual intercourse by the defendant inserting his finger into the complainant's anus.[63] The complainant's evidence was challenged in cross-examination on the basis of apparent discrepancies in his contextual account (such as the furnishings of the defendant's house and the movie watched on the night), but was partly corroborated by evidence from a witness who customarily allowed the defendant to look after her house, which had furnishings similar to those described by the complainant.[64]

The strong case involved three counts, two of indecency on separate occasions a week apart, by the defendant forcing the complainant to masturbate him, and one of sexual intercourse (on the second occasion) by forcing his penis into the complainant's mouth.[65] In this case, the complainant's evidence was corroborated by evidence of his mother and his best school-friend confirming a change in

56 These figures have been rounded for the purpose of this discussion. For the exact figures, see ibid 73–4.
57 Ibid.
58 Ibid 74.
59 Ibid 72.
60 Ibid 78, 84–88.
61 Ibid 76.
62 Ibid.
63 Ibid 84–6.
64 Ibid 86.
65 Ibid 84–5.

his behaviour around the time of the alleged incidents, which the complainant confided to his friend was due to the defendant doing 'sexual things to his willy'.[66]

The weak case involved a single count of indecency.[67] While the complainant was cleaning the defendant's pool after a huge thunderstorm, the defendant allegedly pushed him playfully into the pool, and then convinced him to disrobe and have a shower.[68] The defendant then snuck into the bathroom while the complainant was drying himself, started drying him, and then reached over his shoulder and grabbed and stroked his penis, while rubbing up against him with an erect penis.[69] The incident was allegedly preceded by a number of vaguely sexual incidents engineered by the defendant, including a surprising story that on the first day of football training, in full view of parents and other boys, the defendant grabbed the complainant's crotch through his new football shorts.[70] However, the complainant's evidence was undermined by contextual discrepancies raised in cross-examination, such as the fact that there was no rain during the period alleged and that neither the defendant's house, nor any of the premises in the defendant's street at the time, had pools (although none of these suggestions was ever formally proven by the defence).[71]

In each of the cases, the defendant gave uncorroborated evidence denying the pertinent facts, and his counsel alleged fabrication by the complainants with various motives.[72] There were no witnesses to any of the charged acts, so all three cases came down to issues of credit. Each complainant gave evidence that they did not know each other, and the possibility of collusion was not taken up by defence counsel.

## E *The Trial Configurations*

In order to accommodate a wide variety of comparisons, ten different trial configurations were employed. The moderate-strength case formed the basis for the basic separate trial and was treated as the focal case.[73] In Trial 1, this was accompanied by standard jury directions.[74] In Trial 2, the evidence was

66 Jane Goodman-Delahunty, Annie Cossins and Natalie Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study — Trial Scripts* (May 2016) Royal Commission into Institutional Responses to Child Sexual Abuse <https://www.childabuseroyalcommission.gov.au/getattachment/ad917bcb-4f20-4ee0-ab50-27b2f56a70b6/Jury-reasoning-trial-scripts> 107. In some trial configurations, the corroborating evidence of the mother and school friend was omitted.

67 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 84.

68 Ibid.

69 Ibid.

70 Ibid.

71 Ibid.

72 Ibid 86–8; Goodman-Delahunty, Cossins and Martschuk, *Trial Scripts*, above n 66, 50, 54–8, 61–4, 68, 81–2.

73 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 76.

74 Ibid.

supplemented by so-called 'relationship' evidence involving alleged grooming behaviours, which included getting the complainant to parade around and pose either scantily clad or nude, like a model, while the defendant looked on and occasionally took photos.[75] In Trial 3, this was supplemented by specific judicial directions cautioning jurors to use the relationship evidence only to provide context, not tendency, and not as a substitute for the charged facts.[76] In Trial 4, this was supplemented further by a question trail designed to step the jury through the factual issues it needed to resolve.[77] Question trails have been found to assist jury reasoning in complex cases,[78] so their effects were a secondary target of investigation in the JT Study. Trials 5 and 6 involved separate trials of the focal case supported by evidence from the weak and strong cases admitted as tendency evidence.[79] Trial 5 employed the standard judicial direction while Trial 6 used a direction specifically cautioning against improper use of the tendency evidence.[80]

Trials 7 to 10 were joint trials of six counts with the three complainants giving the same weak, moderately strong and strong evidence as they gave in the separate trials.[81] From the perspective of the focal case, the additional evidence from other complainants provided the same tendency evidence as in Trials 5 and 6. In Trials 7, 8 and 10, there were six prosecution witnesses — the three complainants, one corroborator in the moderate case and two in the strong case.[82] Trial 10 had the standard judicial directions, while in Trials 7 and 8 these were supplemented by tendency directions.[83] Trial 8 was distinguished by also having a question trail.[84] Trial 9 had only four prosecution witnesses (three complainants and the corroborator of the moderate case), with both standard and tendency directions unsupported by a question trail.[85]

Table 1 below summarises the mock trials conducted and their configurations. As will appear, the number of juries and jurors for each type of trial is important in interpreting the quantitative results, since the sample sizes affect statistical power. Outcomes were analysed not only in terms of verdict, but also in terms of more specific aspects of reasoning assessed through the post-trial questionnaire and an analysis of transcripts of the jury deliberations.

75 Ibid; Goodman-Delahunty, Cossins and Martschuk, *Trial Scripts*, above n 66, 25, 28, 30, 35.

76 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 76; Goodman-Delahunty, Cossins and Martschuk, *Trial Scripts*, above n 66, 44–5.

77 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 76, 330.

78 See Catriona McKay, Mark Nolan and Michael Smithson, 'Effectiveness of Question Trails as Jury Decision Aids: the Jury's Still Out' (2014) 21 *Psychiatry, Psychology and Law* 492.

79 Ibid 76.

80 Goodman-Delahunty, Cossins and Martschuk, *Trial Scripts*, above n 66, 84–6.

81 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 76.

82 Goodman-Delahunty, Cossins and Martschuk, *Trial Scripts*, above n 66, 92–115.

83 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 76–7.

84 Ibid 77.

85 Ibid 179.

**Table 1[86]: Summary of mock trials by type and features**

| Trial (ID) | No. of juries | No. of jurors | Trial type | No. of accusers | Evidence types | Prosecution witnesses | Judicial directions |
|---|---|---|---|---|---|---|---|
| 1 | 9 | 105 | Separate | 1 | Basic | 2 | Standard |
| 2 | 12 | 135 | Separate | 1 | Relationship | 2 | Standard |
| 3 | 9 | 103 | Separate | 1 | Relationship | 2 | Standard + relationship |
| 4 | 10 | 107 | Separate | 1 | Relationship | 2 | Standard + relationship + question trail |
| 5 | 8 | 85 | Separate | 3 | Tendency | 4 | Standard |
| 6 | 9 | 112 | Separate | 3 | Tendency | 4 | Standard + tendency |
| 7 | 8 | 93 | Joint | 3 | Tendency | 6 | Standard + tendency |
| 8 | 8 | 100 | Joint | 3 | Tendency | 6 | Standard + tendency + question trail |
| 9 | 9 | 108 | Joint | 3 | Tendency | 4 | Standard + tendency |
| 10 | 8 | 83 | Joint | 3 | Tendency | 6 | Standard |

Results of the differing trial types were assessed by statistical comparisons in which 'significance' is judged by a conventional standard based on the unlikelihood of the result occurring by chance.[87] While scientists do not accept the concept of absolute proof, they recognise experimental support for an effect if it is unlikely to have occurred by chance, with the conventional threshold usually being arbitrarily set at five per cent. If a difference between two conditions (eg types of trial) has less than a five per cent chance of occurring by chance (expressed as $p < .05$),[88] then scientists conventionally accept that it is likely to have occurred due to a functional difference between the two conditions.

## IV RESULTS AND INTERPRETATION

### A *Comparison of Conviction Rates*

Verdicts were recorded for the jury groups immediately after deliberations, then later on an individual basis in the post-trial questionnaire.[89] Consistent with prior research, group judgments were generally more lenient overall than individual judgments.[90] Some theories to explain this are canvassed at pages 66–7 of the

86 Data taken from Table 2, ibid 77.
87 Ibid 20.
88 'p' stands for 'probability'.
89 Ibid 81, 335.
90 Ibid 96.

Report.[91] The JT Study contributes to the body of knowledge in this area by finding this group-individual difference even when the individual judgment is tested after, rather than before, the group deliberations.

## B *The Effect of Relationship Evidence*

Trials without tendency evidence were used to test the effect of the relationship evidence, relationship direction and a question trail. The pattern of results in these cases was quite curious. The JT researchers report (at page 94) that '[c]onviction rates for both the non-penetrative and penetrative offences against the focal complainant in the relationship evidence trial … were significantly higher than those in the basic separate trial', but that is at odds with what is reported at pages 112 and 113. Table 2 below, which shows the results of the basic trial and the various relationship trials, suggests that it depends on which comparison one is performing.

**Table 2[92]: Comparison of verdicts in trials of the focal case without tendency evidence**

| Trial type | Jury direction | Group guilty verdicts (%) | | Individual guilty verdicts (%) | |
|---|---|---|---|---|---|
| | | Indecency | Intercourse | Indecency | Intercourse |
| Standard, without relationship evidence (Trial 1) | Standard | 11.1 | 0.0 | 19.0 | 9.5 |
| Standard + relationship evidence (Trial 2) | Standard | 8.3 | 0.0 | 22.2 | 19.3 |
| Standard + relationship evidence (Trial 3) | Standard + relationship | 33.3 | 33.3 | 68.9 | 69.9 |
| Standard + relationship evidence (Trial 4) | Standard + relationship + question trail | 10.0 | 0.0 | 23.8 | 19.2 |

For the most serious offence, the conviction rate was zero for three of these trial types. The addition of relationship evidence in Trials 2 and 4 had no significant effect on group verdicts, although it appeared to reduce individual jurors' reluctance to convict on the more serious offence.[93] When the relationship

91 Ibid 96.

92 These figures are taken from Table 4 of Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 96.

93 The statistic used to assess these differences is the chi-square ($\chi 2$) statistic. The difference in individual verdicts in Trials 1 and 2 for the more serious offence was statistically significant: $\chi 2 = 4.39$, $p = .036$.

evidence was accompanied by a specific judicial direction on the use of that evidence, the conviction rate rose substantially on both charges, but only when it was *not* also accompanied by a question trail. The JT researchers report that the difference between Trials 2 and 3 was not statistically significant,[94] which is true of jury verdicts because of the small sample size, but at the juror level, the differences were highly significant.[95]

They also suggest, based on results of the post-trial questionnaire, that the relationship direction may have confused the jurors,[96] although elsewhere they argue that the relationship evidence itself may have had that effect by increasing the rate of hung juries.[97] The effect of the judicial direction will be considered in more detail later.

In view of the lack of uniformity in the results of relationship trials and the potentially confusing effect of the judicial direction, findings in the Report with respect to relationship trials in general must be read with some caution.

### C *Trials with Tendency Evidence*

The trials with tendency evidence included both separate trials of the focal case in which the tendency evidence of the other complainants was included solely as corroboration, and joint trials in which all allegations were the subject of charges. The conviction rates in trials with tendency evidence were significantly higher than in trials without such evidence.[98] These conviction rates appear in Table 3 below. As the JT researchers observe, this does not of itself indicate that any illogical or impermissible reasoning took place, since the tendency evidence may have been logically persuasive.[99] This issue was pursued further in the analysis of jury deliberation and reasons, which will be dealt with later.

94 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 210.

95 For the minor offence, $\chi 2 = 52.18$, $p < .0001$; for the major offence, $\chi 2 = 61.87$, $p < .0001$. These calculations are based on reconstructing the raw data from the percentage figures.

96 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 204, 209 (Figure 11).

97 Ibid 113.

98 Ibid 96.

99 Ibid 114, 126–7.

**Table 3: Comparison of verdicts in the focal case in trials with tendency evidence[100]**

| Trial type | Jury direction | Group guilty verdicts (%) | | Individual guilty verdicts (%) | |
|---|---|---|---|---|---|
| | | Indecency | Intercourse | Indecency | Intercourse |
| Standard + tendency evidence (Trial 5) 4 witnesses | Standard | 62.5 | 62.5 | 71.4 | 66.7 |
| Standard + tendency evidence (Trial 6) 4 witnesses | Standard + tendency | 55.6 | 55.6 | 75.5 | 76.4 |
| Joint trial (tendency) (Trial 10) 6 witnesses | Standard | 100.0 | 75.0 | 93.8 | 84.0 |
| Joint trial (tendency) (Trial 9) 4 witnesses | Standard + tendency | 100.0 | 100.0 | 98.1 | 95.3 |
| Joint trial (tendency) (Trial 7) 6 witnesses | Standard + tendency | 75.0 | 75.0 | 88.2 | 88.0 |
| Joint trial (tendency) (Trial 8) 6 witnesses | Standard + tendency + question trail | 87.5 | 87.5 | 90.9 | 88.9 |

Although it was reported that there was no statistically significant difference between verdicts in the joint and separate tendency trials, leading to a conclusion that there was no joinder effect,[101] the comparison referred to does not test for a joinder effect as defined in the JT Study, but rather for a joinder effect per se, defined earlier herein. Furthermore, even with respect to the joinder effect per se, the tabulated results show divergences which most jurists would regard as having practical significance.

Table 3 allows a comparison of conviction rates in the moderate-strength focal case in separate and joint trials having substantially the same tendency evidence. The nearest to a perfect comparison is between Trials 6 and 9, which contained identical evidence and only differed on whether the weak and strong complaints were also the subject of joined charges. The joinder in that situation led to a perfect conviction rate, almost doubling the rate in the same trial without the joinder of charges. Other comparisons are imperfect due to the differences in the number of witnesses corroborating the strong complaint, and (in Trial 8) by the use of a question trail, but in all instances the joinder of charges in almost

100 Figures taken from ibid 96 (Table 4).

101 Ibid 24, 94.

identical cases led to an increase in the conviction rate which in practical terms would be regarded as significant, and disturbing. Neither the specific judicial direction on the use of tendency evidence nor the question trail seemed to have any consistent effect.[102]

In Table 14 of the Report,[103] the JT researchers combined Trials 5 and 6 to represent separate tendency trials and Trials 7 and 10 to represent the joint trials. The relevant figures are reflected in Table 4 below.

**Table 4: Guilty verdicts in the focal case for juries and individual jurors in joint and tendency trials**

| | Trial type | Sample | Non-penetrative offence | | Penetrative offence | |
|---|---|---|---|---|---|---|
| | | | % guilty | no. guilty* | % guilty | no. guilty* |
| Juries | Joint | 16 | 87.5 | 14 | 75.0 | 12 |
| | Tendency | 17 | 58.8 | 10 | 58.8 | 10 |
| Jurors | Joint | 176 | 90.8 | 160 | 86.1 | 152 |
| | Tendency | 197 | 73.7 | 145 | 72.2 | 142 |

* These raw figures are reconstructed from the percentages in Table 14.

The following passage reporting on these figures is problematic for a number of reasons:

> The conviction rate by juries in the joint trial was slightly higher than in the separate trial with tendency evidence: 87.5 per cent and 75 per cent versus 58.8 per cent ... Chi-square analyses of jury verdicts — convictions versus acquittals plus hung juries — revealed no difference in conviction rates in the tendency evidence trial compared to the joint trial.[104]

Most people would not regard a difference of 87.5 per cent, or even 75 per cent, versus 58.8 per cent, as 'slightly higher', especially if one takes into account what is being measured — conviction of a serious crime. Nor does chi-square analysis entirely support the assertion of 'no difference' at the jury level.[105] For the minor offence, the difference between 14 guilty verdicts (out of 16 trials) versus 10 guilty verdicts (out of 17 trials) had less than a 12 per cent likelihood of arising by chance.[106] When the same jurors' individual verdicts are subjected to chi-square analysis, the statistical sample becomes much larger, and the less pronounced percentage differences are found to be highly significant statistically.[107]

102 This was confirmed by later more refined analysis in ibid 213–14, 223.

103 Ibid 184.

104 Ibid.

105 Clearly, the JT researchers are referring to lack of statistical significance, but lack of statistical significance does not justify an assertion of 'no difference'.

106 The figures footnoted in the Report at footnote 338 *do* show a significant difference ($p = .023$), but it appears that the wrong figures have been quoted: Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 184. Based on a reconstruction of the raw data from the quoted percentages, the correct figures are: Minor offence: $\chi 2 = 3.417$, $p = .065$, Fisher's exact = .118; Major offence: $\chi 2 = 0.971$, $p = .325$, Fisher's exact = .465. With these small samples, the standard p-calculation is regarded as unreliable, so Fisher's exact calculation is preferred as the measure of probability.

107 Jurors: Minor offence: $\chi 2 = 18.67$, $p < .0001$; Major offence: $\chi 2 = 11.358$, $p < .001$.

However, when analysing individual juror verdicts, the JT researchers employed a different statistical methodology known as regression.[108] Regression is a technique for estimating the *independent* contributions of predictor variables to an outcome. Its use is sometimes deprecated when the predictors are highly *interdependent*, because it masks common or shared effects by arbitrarily assigning such effects to one predictor or the other.[109] In this case, the JT researchers tested the contributions of prior child sexual abuse knowledge (derived from the pre-trial questionnaire), convincingness of the complainant (derived from the post-trial questionnaire), and trial type (either separate tendency trial or joint) to individual juror verdicts. They found that convincingness predicted individual verdicts on all counts, but trial type only predicted verdicts on the strong counts. On the moderately strong focal case, the likelihood of an effect for trial type was 92.1 per cent for the minor offence and 87 per cent for the major offence, which falls short of the arbitrary statistical threshold of 95 per cent, but hardly warrants the JT researchers' conclusion of 'no effect'.[110] However, the main problem with such an analysis is that it assumes that perceptions of the complainant's convincingness are independent of the type of trial, whereas the figures quoted at page 107 of the Report suggest a strong relationship, with the complainant's convincingness rated much higher in the joint trial than in the tendency trials with the same or similar evidence. In fact, these figures suggest that complainant convincingness is itself subject to a statistically significant joinder effect per se.[111]

## D *Other Evidentiary Evaluations*

The JT researchers also examined individual jurors' evaluations of the case through post-trial questions about issues relating to possible guilt, such as the factual culpability of the defendant and the credibility of the complainant.[112] Most of those questions were rated by jurors on scales with ranges like '[n]ot at all' to '[v]ery much' or '[s]trongly disagree' to '[s]trongly agree'.[113] Factual culpability was really just an assessment of the likelihood of guilt on a seven-item scale rather than a binary verdict.[114] The scores on these measures generally mirrored the binary judgments of guilt in the sense that higher conviction rates correlated with higher average evaluations of factors pointing to guilt and lower evaluations

108 Goodman-Delahunty, Cossins and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 185.

109 See Donald E Farrar and Robert R Glauber, 'Multicollinearity in Regression Analysis: The Problem Revisited' (1967) 49 *Review of Economics and Statistics* 92; Michael A Poole and Patrick N O'Farrell, 'The Assumptions of the Linear Regression Model' (1971) 52 *Transactions of the Institute of British Geographers* 145, 148; Andy Field, *Discovering Statistics Using IBM SPSS Statistics* (SAGE Publications, 4th ed, 2013) 324–5.

110 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 185 n 342: for the minor offence, $p = .079$; for the major offence, $p = .130$.

111 For calculations, see below n 137.

112 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 336–40.

113 Ibid.

114 Ibid 16–17.

of factors pointing to innocence.[115] However, in drawing conclusions from these measures, the JT researchers again severely narrowed the conception of a joinder effect by treating these variables as independent from joinder. The following passage again seems problematic:

> Rather, the perceived culpability of the defendant and the credibility of the focal complainant increased in response to independent sources of evidence, not more evidence or the type of trial. As more independent sources of evidence were introduced to support the focal complainant's account, his credibility increased and his evidence was accorded more weight.[116]

The first conclusion, that culpability and credibility judgments increased due to independent sources rather than 'more evidence', does not appear to be warranted by either the design of the JT Study or the results. The study design only added 'more evidence' by way of:

(a) relationship evidence (from the complainant);

(b) tendency evidence (from independent sources); and

(c) independent corroboration of the complainant in the strong case (which evidence was omitted in the separate tendency Trials 5 and 6 and the joint Trial 9).

The non-independent relationship evidence *did* increase assessments of the defendant's culpability and the focal complainant's credibility, though not as much as in tendency trials.[117] There was no other non-independent 'more-evidence' condition for comparison. Furthermore, as we shall see later, the addition of two independent corroborating witnesses in the strong case actually *reduced* guilty verdicts and evaluations aligned with guilt.[118] However, the relevance of the independence of witnesses was supported by subsequent qualitative analysis of jury deliberations, since in trials with no tendency evidence, jurors were more likely to express the view that it was one person's word against another.[119]

The second conclusion, that the increased ratings of defendant culpability and complainant credibility were due to independent sources of evidence rather than type of trial, seems to narrow the concept of type of trial to a point where it vanishes entirely. It treats type of trial as something separate from its consequential impact on the evidence admitted. No jurist would limit joinder effects in this way, and the JT Study itself does not purport do so. Rather, joinder is postulated to have adverse effects of character prejudice, irrational accumulation or conflation through being the very mechanism that introduces the evidence from independent sources. The JT Study shows that the introduction of such evidence does increase guilty verdicts significantly. The question remains whether that increase is due to irrational prejudice or rational integration of such evidence into the corpus.

115 See, eg, ibid 107.
116 Ibid 94. See also at 107.
117 Ibid 106, 110.
118 Ibid 194, 197.
119 Ibid 107.

Given the above, the following conclusion is unwarranted:

> The factual culpability ratings were higher in the trials with tendency evidence, that is, in the tendency evidence trials and joint trials. This finding indicated that juries' perceptions of the defendant's guilt responded to the strength of the inculpatory evidence, and not to the type of trial per se. This was also contrary to what many judges and practitioners anticipate juries will do in a joint trial.[120]

## E *Effects on Recall*

Post-trial tests also assessed factual recall by a series of six multiple-choice questions.[121] Recall was significantly undermined by the addition of tendency evidence in both the separate and joint trial modes.[122] Since this effect was not found in the relationship trial, which had no extra witnesses and only a modicum of additional evidence, the JT researchers' conclusion that recall is affected by the complexity of the trial and the number of witnesses is well founded. However, they again attempted to sever this effect from the effects of joinder:

> Notably, the potential for factual inaccuracy or confusion was not based on the type of trial per se, and was not the result of joinder. Rather, the similarity in the allegations by the three males led to the additional error, as the increase in the error rate was similar, irrespective of whether this information was presented in a separate trial, as tendency evidence, or in a joint trial as cross-admissible tendency evidence.[123]

Again, this passage reflects an approach of joinder per se. Nobody would suggest that joint trials increase the potential for errors of recall other than by increasing the complexity (or volume) of the evidence, so complexity and trial type are not independent variables. The study findings support the hypothesis that increasing the complexity of a trial by allowing joinder of multiple charges is likely to have some adverse effect on juror recall.[124] Subsequent analysis of jury deliberations supports this conclusion, and also a joinder effect per se, since factual error rates were noticeably higher in joint trials compared to separate tendency trials.[125]

The application of these findings to the real world is another issue. Analysis of the transcripts of jury deliberations suggested that many individual recall errors were corrected by the group, but group verdicts did significantly correlate with factual error rates, with the more accurate jury groups favouring acquittal.[126] However, the longest simulated trial lasted only 110 minutes,[127] and deliberations began immediately thereafter, so the opportunity for memory interference, conflation and degradation was limited. In trials that might last days with

120 Ibid 98.
121 Ibid 99–101.
122 Ibid 100.
123 Ibid 100.
124 Ibid 224.
125 Ibid 115–6 (Table 7).
126 Ibid 100–1.
127 Ibid 116, 254.

deliberations taking place long after a particular witness was heard, the scope for memory error is greater and the capacity of the group to correct it somewhat less. The JT researchers, on the other hand, suggest that the abbreviated trial and deliberation period may have *increased* cognitive load (a suggestion which is highly debatable, and contrary to their own definition of cognitive load theory)[128] and thereby made jurors 'more vulnerable to heuristic reasoning, confusion and errors'.[129] Elsewhere (though not in the 'Limitations of the study' section of the Report), they concede the increased potential for confusion in longer real life trials through 'trial length, juror fatigue, juror disinterest, changing levels of concentration and trial complexity'.[130]

### F *Effects of Joinder Per Se on Evidentiary Evaluations*

While juror evaluations on various issues generally aligned consistently with guilty verdicts, the differences between tendency trials and joint trials on similar or identical evidence were also consistently found in these evaluations. On the following measures, separate trials with tendency evidence produced evaluations more favourable to the defendant than joint trials on the same or similar evidence:[131]

- Factual culpability of the defendant for both counts in the focal complaint;[132]
- Factual culpability of the defendant for all three counts in the strong complaint;[133]
- Perception of whether the defendant had a sexual interest in boys;[134]
- Perception of the criminal intent of the defendant;[135]
- Credibility of the focal complainant;[136]
- Convincingness of the focal complainant;[137]

128 Ibid 15.

129 Ibid 268.

130 Ibid 33.

131 For factual culpability, figures were available to compare Trial 5 v Trial 10 (tendency and joint trials with identical evidence plus the standard judicial direction) and Trial 6 v Trial 9 (tendency and joint trials with identical evidence plus both standard and tendency directions) — ibid 96. For the other measures, figures were only available for Trials 5 and 10 — at 98.

132 Ibid 110.

133 Ibid.

134 Ibid 102.

135 Ibid 102–3. Perception of criminal intent was measured as a composite of responses on questions about whether the defendant 'abused the trust of others', 'abused his position as a coach', 'was responsible for what happened to him' and 'was a risk to other boys'.

136 Ibid 106. Credibility was assessed by standardised tests of *Poise*, which assesses confidence, emotional control and anxiety management, and *Communication Style*, which assesses verbal and non-verbal communication associated with witness trustworthiness, likeability and knowledge, both from the *Observed Witness Efficacy Scale* published in Robert J Cramer et al, 'The Observed Witness Efficacy Scale: A Measure of Effective Testimony Skills' (2013) 43 *Journal of Applied Social Psychology* 1696.

137 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 107.

Statistical significance was not reported for most of these comparisons, but on the measures of factual culpability and complainant's convincingness, sufficient data were reported to calculate that the differences were often highly significant.[138] These measures show a trend similar to that for verdicts, that joint trials lead to higher assessments of guilt than tendency trials on the same or similar evidence. The factors that bucked this trend were:

(a) perception of victim blame, which was higher in the joint trial compared to separate tendency trials, despite the fact that this might point towards innocence rather than guilt;[139]

(b) assessments of the *defendant's* convincingness, where only the relationship evidence appeared to be influential;[140] and

(c) factual culpability of the defendant in the weak case.[141]

## G *The Relationship Between Trial Type and Evidentiary Evaluations*

On the evaluative factors unfavourable to the defendant (other than the three exceptions noted above), the results matched this pattern:

*Basic trial < relationship trial < separate tendency trial < joint trial* [142]

The JT researchers argue that the addition of extra evidence independent of the focal complainant rationally increased assessments of guilt and culpability,[143] but this does not explain why:

(a) verdicts and evaluations tending towards guilt were higher in joint trials compared to tendency trials on the same or similar evidence;[144] and

(b) (as we shall see below), both verdicts and assessments of defendant culpability *decreased* when two additional prosecution witnesses were added in the joint trials.[145]

138 Ibid. For factual culpability, Table 4, at 96, provides sufficient information to perform *t*-tests which compare the means for different types of trial. Although Table 4 does not make it clear, Table 5, at 110, shows that the bracketed figures are the standard deviations required for these calculations: eg Trial 6 v Trial 9 Focal complaint Minor offence: $t\,(218) = -3.48$, $p < .001$; Major offence: $t\,(218) = -3.83$, $p < .001$; Trial 5 v Trial 10 Minor offence: $t\,(166) = -2.33$, $p = .021$; Major offence: $t\,(166) = -1.79$, $p = .075$. For complainant's convincingness, standard deviations are at 107, footnote 187, for Trial 5 and Trial 10 (see also footnote 184): $t\,(166) = -2.996$, $p = .003$.

139 Ibid 104.

140 Ibid 239–40.

141 Ibid 110.

142 See above nn 132–7.

143 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 25.

144 Ibid 101–10.

145 Ibid 194, 197.

Neither of these findings is consistent with the systematic, rational use of the additional independent evidence tendered in support of the prosecution case.

## H *The Effect of Judicial Directions and Question Trails*

As mentioned earlier, the tendency direction appeared from quantitative analysis to have no systematic effect. Qualitative analysis of transcripts suggested that the jurors struggled with the notion of using the tendency evidence, in effect retrospectively, to bolster the weak case which occurred first in time.[146] Many juries either ignored or misunderstood the lengthy tendency direction, and only one used it correctly to reach a guilty verdict on the weak count.[147] Jurors were more likely to rate the judge's direction as confusing when the tendency direction was included, but it was also more likely to change their minds about the verdict.[148]

On the other hand, the relationship direction merely seemed to confuse. This conclusion was drawn from an analysis of post-trial questions testing the effect of the direction, which showed that it made:[149]

- the judge's instructions more confusing;
- the charges harder to understand;
- the facts harder to recall;
- the tasks of assessing witness credibility and applying the law to the facts more difficult and effortful; and
- the tasks of weighing the evidence and assessing the prosecution's case more difficult.

Jurors also made little mention of this judicial direction in their deliberations.[150]

These adverse effects appeared to be offset by the addition of a question trail, but the following conclusion of the JT researchers is too broad:

> Separate analyses in relationship evidence trials showed that with the aid of a question trail, juries rated the defendant as significantly less factually culpable, and accordingly, the conviction rate for both penetrative and non-penetrative offences declined.[151]

This conclusion was based on a comparison of Trial 3 (with the relationship direction) and Trial 4 (with both the relationship direction and a question trail).[152] As Table 2 shows, this was only true when the question trail was added to the

146 Ibid 148–9. See analysis of deliberations for Jury 75 (at 159), who convicted on the weak count, and the acquittals by Jury 83 (at 165), Jury 54 (at 173–4), and Jury 63 (at 174–5).

147 Ibid 177.

148 Ibid 214.

149 Ibid 204, 209.

150 Ibid 210.

151 Ibid 222.

152 Ibid 223 n 464.

confusing relationship direction. The question trail had no apparent effect when compared to the relationship trial without the relationship direction (Trial 2).

When a question trail was used in the joint trial with tendency rather than relationship evidence,[153] no significant effect was found on verdicts, factual culpability assessments, recall or cognitive effort, except that jurors reported significantly more difficulty in understanding the charges with the question trail than without it.[154]

### I *Comparison of Verdicts on Minor and Major Offences*

The verdicts also reveal another surprising discrepancy. Although the two counts in the focal case arose out of the same transaction, and the only apparent issue was general credit, there were significant differences between the verdicts on each count with the same complainant. This difference did not appear to be influenced by the type of trial.[155] In general, jurors were more reticent to convict on the penetrative offence, and this difference was mirrored in post-trial ratings of the defendant's culpability for each count.[156] The JT researchers talked up this result as showing that the jurors were assessing each count separately and rationally,[157] a practice that would appear virtuous if jurors were assessing discrete cases, but when the counts are both part of a single transaction, and the only apparent issue is global credit, one would expect the same verdict on both counts. The JT researchers attribute the discrepancy to some unspecified differential in processing of the evidence about the penetrative and non-penetrative offences,[158] but an alternative explanation is that the jurors were unconsciously adjusting the strength of evidence required according to the seriousness of the charge, an approach which would find some indirect support from cases like *Briginshaw v Briginshaw*[159] and *Neat Holdings Pty Ltd v Karajan Holdings Pty Ltd*.[160]

## V ANALYSIS OF PRIMARY RESEARCH QUESTIONS

The post-verdict questionnaire was not merely used for discrete analysis of trial effects. Some questions were also used in combination with qualitative analysis of deliberations to test the specific theories of character prejudice, accumulation and conflation. These analyses are best understood in that context.

153 Trial 8 v Trial 7.

154 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 230.

155 Ibid 112, 252.

156 Ibid 110, 113.

157 Ibid 94.

158 Ibid 109, 113.

159 (1938) 60 CLR 336.

160 (1992) 110 ALR 449.

### A *Character Prejudice*

Since increased conviction rates in trials with tendency evidence shed no light on whether the verdicts were based on permissible or impermissible reasoning, the JT researchers sought to detect such prejudice through other quantitative measures and a qualitative analysis of jury deliberations. Quantitatively, they argued that if character prejudice were present, they would expect to see in tendency trials (both separate and joint):[161]

(a) decreased perception of the defendant's convincingness;

(b) decreased perception of cognitive effort due to the use of effortless heuristic reasoning;

(c) undifferentiated ratings of the defendant's criminal intent; and

(d) undifferentiated ratings of the defendant's culpability.

With the greatest respect, these predictions seem shaky at best.

With respect to prediction (a), the defendant's convincingness is likely to be an assessment of his presentation in the witness box, which may or may not be tainted by extrinsically generated reservations about his character. Jurors may even have interpreted it as an assessment of his *performance*, rather than whether he actually convinced them.[162] Furthermore, if ratings of convincingness *were* tainted by extrinsic judgments of character, one might also expect them to be tainted by adverse judgments based on permissible reasoning about tendency, so the prediction is arguably incapable of distinguishing character prejudice from legitimate reasoning. The JT researchers reported that 'jurors rated the defendant equally convincing in the basic separate trial, the tendency evidence trial and the joint trial'.[163]

Prediction (b) assumes that effortless heuristics displace parallel conscious reasoning, leading to an overall reduction in cognitive effort. However, the group jury context required participants to reason consciously, and to rationalise their views to fellow jurors, which rules out effortless processing and may even be more effortful for an irrationally prejudiced juror.

The final two predictions are even more obscure. Given that tendency evidence is supposed to corroborate the complainant and increase the conviction rate, one would expect measures analogous to guilt, such as criminal intent and factual culpability, to be similarly affected. As reported earlier, these ratings were higher in tendency cases. However, the JT researchers seem to be suggesting that character prejudice, if present, would completely override any other assessment

161 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 259–60.

162 Jurors were simply presented with the proposition: 'The accused Mark Booth was convincing', and asked to rate their agreement with that proposition on a scale from 1 ('strongly disagree') to 7 ('strongly agree'): ibid 340 (Appendix L).

163 Ibid 260.

of the strength of the case, such that findings on criminal intent and culpability would be identical for weak, moderate and strong cases:

> Comparing the factual culpability ratings for the conduct alleged by each of the three witnesses or complainants gave a further indication that juries did not reason globally and return parallel ratings about the defendant on measures of culpability in trials with tendency evidence, but made appropriate distinctions between the evidence and counts.[164]

The prediction of such undifferentiated ratings seems unwarranted. The fact that jurors draw distinctions between evidence and counts in assessing culpability and criminal intent does not rule out any underlying influence of prejudice in their assessments.

The JT researchers acknowledge that these quantitative findings on character prejudice are insecure, on the grounds that jurors' views could have been influenced by permissible rather than prejudicial reasoning, and in any event post-verdict self-reporting may not accurately reflect the cognitive process during deliberations.[165] The analysis of the transcribed deliberations was therefore fundamental to identifying whether the differences in guilt evaluations were the result of permissible reasoning or impermissible character prejudice.

## B *Analysis of Transcripts of Deliberations*

The analysis of the deliberations involved both quantitative studies, based on the presence or absence of coded features, and qualitative analysis of transcripts of jury deliberations.

For the quantitative stage, trained research assistants were presented with the transcripts of the recorded deliberations and asked to code them for various attributes:[166]

- juries' understanding of the evidence;
- juries' understanding of the judicial directions; and
- the presence of factual errors, unfair prejudice against the defendant, and any verdicts motivated by inter-case conflation of the evidence, accumulation prejudice, and/or character prejudice.

Appropriate preliminary steps were taken to ensure consistency in coding between different research assistants (called inter-rate reliability). The codings listed for unfair prejudice are expressed as follows:[167]

- reason emotional reaction to case;
- convict for emotional reaction;

164 Ibid.
165 Ibid 260–1.
166 Ibid 82.
167 Ibid 346.

- evidence logically unconnected to reason towards guilty verdict;
- convict logically unconnected;
- reason lower threshold than 'beyond reasonable doubt' should be used;
- convict lower threshold than 'beyond reasonable doubt'.

These codings were explained as follows: 'Jury deliberations were coded to assess whether they demonstrated impermissible reasoning in one of three ways: applying a lower standard of proof; reasoning emotionally, based on the evidence; or reasoning illogically about the evidence.'[168]

There is no other explanation of what the researchers were looking for, except that illogicality was equated with reasoning that was 'unrelated to evidence'.[169] What is clear is that these methods require the impermissible reasoning to be explicitly manifested, as suggested by this passage:

> Only two juries out of 90 contained a juror who indicated that they were basing their verdict on a lower standard of proof than the criminal standard of proof. Both of these jurors indicated that they were making their assessments of the evidence 'on the balance', that is, on a standard closer to the balance of probabilities.[170]

However, the most obvious objection with respect to unfair prejudice is that these codings are not specifically directed to the definitions of that prejudice — a conclusion which is confirmed by the fact that the JT researchers only found impermissible reasoning in cases that did *not* include tendency evidence.[171] With respect to logicality, judges have often said that reasoning based on propensity is quite logical, but the concern is that a jury will attach too much weight to it.[172] Similarly, reasoning based on the volume of evidence may well be perfectly logical. There is no explanation of how the JT researchers distinguished, simply through an analysis of transcripts, the difference in degree between valid reasoning about propensity or accumulation of evidence and impermissible reasoning in which the tendency evidence is irrationally over-valued. Furthermore, although the Report notes the Australian Law Reform Commission's suggestion that character prejudice could be actuated by emotions of anger or outrage, it also noted that such emotions may act 'consciously or unconsciously'.[173] The overarching impression from these codings is that unfair prejudice was simply equated with manifestly illogical reasoning, overt emotion or explicit easing of the burden of proof. This seems a far cry from the rigour suggested by: '[w]e must delineate and define them with specificity before testing them'.[174]

168 Ibid 120.

169 Ibid 121 (Table 9).

170 Ibid 120.

171 Ibid 121 (Table 9). Given the definitions of prejudice, they could not arise in the non-tendency cases.

172 See, for example, the oft-quoted dictum of Lord Cross in *DPP v Boardman* [1975] AC 421, 456.

173 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 47, citing Law Reform Commission, *Evidence*, Interim Report No 26 (1985) vol 1, 456–8 [800]–[803].

174 Ibid 45.

With respect to conflation, other codings of factual errors and the correction of such errors by the group were more relevant, and this form of prejudice should be more readily detectable in group deliberations.

Analysis based on the codings was of a quantitative nature. The JT researchers also conducted a qualitative thematic analysis of the transcripts to identify instances of character or accumulation prejudice. They adopted a methodology of 'sensemaking' employed in research about social discourse[175] and the Report shows how it was used,[176] but there is nothing to suggest that that methodology can assist in drawing the necessary distinctions of degree between permissible reasoning about propensity and accumulation of evidence and impermissible over-weighting of those factors. What can be said is this: transcripts can only identify prejudice if it is manifested explicitly and with sufficient definition to discriminate it from non-prejudicial manifestations. In relation to character prejudice and accumulation prejudice, that would require words that clearly show that the jury or juror is attaching irrational weight to prior conduct or the accumulation of evidence. There is nothing in the JT Study to demonstrate how such sensitivity could be achieved in a simple textual analysis.

## C *Application of Transcript Analysis to Character Prejudice*

The JT researchers derived this conclusion on character prejudice from the analysis of transcripts:

> A quantitative and qualitative content analysis of jury deliberations revealed that no juries in either the tendency evidence or joint trials used the tendency evidence to conclude that the defendant was guilty because of the number of allegations of prior misconduct he was facing. Furthermore, there was no evidence that emotional reactions to the severity of the allegations — such as a sense of horror regarding the allegations, or a desire to punish the defendant — drove the verdicts. To the contrary, we found evidence that juries were more reluctant to convict the defendant for the counts pertaining to the most serious allegations of sexual intercourse than for indecency.[177]

The first sentence seems to reflect accumulation prejudice rather than character prejudice, and the second deploys emotional reaction as a correlate of character prejudice, which is not an inherent part of the definition. Since the trials were known by the jurors to be simulations only, they were inherently incapable of generating authentic emotional reactions.

175 Ibid 68.
176 Ibid 138–54.
177 Ibid 261.

## D *Conflation of Evidence*

Inter-case conflation of evidence was tested by comparing jurors' recall and ratings of defendant culpability in the different types of trial. Findings on factual errors and recall, and the limitations of those findings, have been discussed earlier.

Both the post-trial questionnaire and coding of transcripts were used to identify factual errors, including errors of conflation. The addition of the tendency evidence did increase factual errors in separate trials, and even more so in joint trials.[178] However, these errors were generally corrected by other jurors, and appeared to have no effect on group verdicts. The JT researchers found no instances of uncorrected errors of conflation.[179]

They also argued that the higher ratings of culpability in the tendency trials compared to non-tendency trials suggested that jurors made systematic inferences logically related to the offences, rather than through conflation,[180] but it is not explained how this rules out conflation. It is not at all clear what effect conflation would have on ratings of culpability. If jurors misattributed strong evidence to the weak case, that would presumably increase culpability measures on the weak case, and perhaps reduce culpability measures on the strong case. On the other hand, if inconsistencies in the weak case were attributed to the moderate or strong cases, then the latter cases would be diminished.

## E *Accumulation Prejudice*

To assess accumulation prejudice, the JT researchers examined whether juries were prone to convict based on the number of charges or the number of witnesses, and whether they were capable of distinguishing between the evidence on each count when reaching their verdicts in joint trials. No evidence was found in either transcripts or post-trial reasons for decisions of jurors explicitly manifesting these forms of prejudice, so the following discussion focuses on the quantitative measures and arguments.

### 1 *Multiple Counts*

This form of accumulation prejudice is based on the theory that 'a defendant will be prejudiced in joint trials because juries are prone to reasoning that the defendant is guilty simply because of the number of charges brought by the prosecution'.[181] One would expect this analysis to centre on a comparison of joint trials and separate trials with the same or similar evidence, since these trials differed only on the number of counts. As we have seen, there were convergent findings that joinder per se led to evaluations on a number of measures which

178 Ibid 116.
179 Ibid 117.
180 Ibid 255.
181 Ibid 257.

were more unfavourable to the defendant.[182] This finding bespeaks some form of accumulation effect due to multiplicity of counts.

The separate trials with tendency evidence did not involve verdicts for the weak and strong cases, but the JT researchers analysed individual assessments of culpability in these cases. In relation to the strong case, juries found the defendant significantly more culpable in the joint trial (where the strong case represented three out of six counts) than in separate trials with the same or similar evidence (where the strong case was merely corroborative).[183] This seems to support the conclusion of a joinder effect per se, but the JT researchers instead argue that it negatives accumulation prejudice, because in the weaker case this difference was not found. Again, the JT researchers assumed that for a prejudice to be found it must overwhelm any other factors whereby jurors distinguish between the weak, moderate and strong cases. In fact, they say this explicitly:

> According to the accumulation prejudice hypothesis, in a joint trial one would expect factual culpability ratings for the defendant would not differ according to allegations of varying evidential strength; that is, that juries would rate the count based on the weakest evidence and the count based on stronger evidence similarly.[184]

The JT researchers also argued for the absence of accumulation prejudice on the basis that the ratings of the complainant's convincingness and credibility were higher in tendency trials than in non-tendency trials. They contended that evidence from independent witnesses of similar criminal conduct enhanced perceptions of the focal complainant's credibility, 'irrespective of whether the defendant was charged with counts pertaining to those individuals'.[185] Since joinder per se also increased ratings of the complainant's convincingness and credibility compared to separate trials on the same or similar evidence,[186] it is difficult to see how these findings negate an effect of accumulation of counts. In addition, credibility was a composite measure of the complainant's performance in the witness box, which was identical in all trials, so any difference on this measure between trial types suggests irrationality.

## 2 *Multiple Witnesses*

This version of accumulation prejudice theorises that juries will infer guilt from the impression created by the number of witnesses, rather than engaging in a systematic review of the evidence for each count in the joint trial.[187] Testing this hypothesis was the reason for having joint trials with either four or six witnesses (Trial 7 v Trial 9). Contrary to logic and expectations, for all six counts, the

182 Ibid 101–10.

183 Ibid. See also at 181–2.

184 Ibid 258.

185 Ibid 256.

186 For convincingness, the difference was statistically significant — see the explanation at n 87.

187 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 258–9.

juror-level guilty verdicts and ratings of the defendant's culpability were higher *without* the extra witnesses, and this was also generally true at the jury level.[188] The JT researchers correctly concluded that this finding discredited the theory of prejudice through accumulation of evidence,[189] but failed to note that it also discredits their conclusion that jurors systematically and rationally incorporated additional evidence into their deliberations. In particular, the additional two witnesses directly corroborated the complainant's evidence in the strong case, and yet the inclusion of those witnesses reduced evaluations of the defendant's culpability and conviction rates on those three counts at both the jury and juror level.[190]

The following conclusion is therefore difficult to support:

> These findings directly controverted the accumulation prejudice hypotheses in relation to multiple witnesses, by indicating that both jurors and juries evaluated the evidence of multiple witnesses based on its probative value, not simply the number of witnesses.[191]

## 3 *Jurors' Stated Reasons for Decision*

Jurors' reasons for decision were also elicited in the post-trial questionnaire through a form that asked two questions — 'What was the main reason for your verdict?' and 'What other factors went into your decision?' with two blank lines provided for each answer.[192] These were then categorised and coded. Three reasons accounted for 87.93 per cent of the guilty verdicts:[193]

(a) consistency of details across the evidence provided by multiple independent witnesses (34.76 per cent);

(b) the strong evidence or credibility of prosecution witnesses (34.15 per cent);

(c) the pattern of grooming behaviour engaged in by the defendant (19.02 per cent).

The JT researchers report that '[n]otably, all three were examples of permissible reasoning in support of a verdict to convict the defendant',[194] but they do not explain how they arrived at that conclusion. Absent collusion, consistency of witnesses is obviously a virtue if the witnesses are all describing the same events, but in this case the 'multiple independent witnesses' were giving evidence about completely different occasions which therefore diverged at the level of detail (despite the researchers' categorisation), so 'consistency' could only be achieved

188 Ibid 194 (Figure 10), 197 (Table 15). At the jury-level, the addition of extra witnesses did not lead to a higher conviction rate for the weak count.

189 Ibid 259.

190 Ibid 194 (Figure 10), 197 (Table 15).

191 Ibid 34.

192 Ibid 284.

193 Ibid 123–4.

194 Ibid 124.

at a more abstract propensity level which may or may not have been overvalued. Similarly, assertions about the strength or credibility of the prosecution evidence tell us nothing about whether that judgment was arrived at by rational means or irrational prejudice. It can be argued that judges, lawyers and lay people all greatly overvalue such similarities.[195]

## VI GENERAL CRITIQUE

Perhaps the most problematic, and disconcerting, aspect of the Report is its confusion of the joinder effect supposedly targeted by the JT Study with the concept of a joinder effect per se. Although the targeted joinder effect is the subject of the Report's definition and introduction, the JT researchers often express conclusions based on the joinder effect per se, giving a very misleading impression of what the JT Study actually established.[196]

Furthermore, despite the fact that the joinder effect per se involves comparing trials with no difference other than joinder of the counts themselves, the JT Study did provide substantial convergent evidence supporting such an effect. At the group jury level, the differences were not statistically significant, but nevertheless substantial in a practical sense.[197] At the juror level, the differences were highly significant.[198]

For reasons stated earlier, quantitative analyses were unable to rule out a targeted joinder effect, so the analysis of transcripts became crucial in testing for impermissible reasoning. This analysis was open to several objections which undermine its findings:

(a) it assumed that prejudice could be detected through the explicit words of jurors in a group deliberation, ignoring the fact that prejudice is likely to be unconscious;

(b) it assumed that prejudice is reflected in manifest divergences from logic, as opposed to undetectable over-weighting of prejudicial evidence;

(c) it failed to detect any illogical reasoning underlying several irrational outcomes; and

(d) it lacked definition in identifying exactly what the researchers were looking for in the transcripted deliberations to distinguish impermissible from permissible reasoning.

195 See Peter M Robinson, 'Prior Convictions, Conduct and Disposition: A Scientific Perspective' (2016) 25 *Griffith Law Review* 197.

196 See, for example, the Executive Summary in Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 24, and the further summary at 94.

197 Ibid 96.

198 Ibid.

The same types of criticism could be levelled at the analysis of jurors' reasons for decision provided in answer to the post-trial questionnaire.

Psychological research shows that judgemental biases are triggered automatically and unconsciously,[199] and judicial observations also imply a more implicit, subversive role:

> The allegation that a number of the accused's relatives died or suffered from arsenic poisoning immediately conjures up a highly suspicious prejudicial atmosphere in which the presumption of innocence tends to be replaced with a presumption of guilt.[200]

It is difficult to see how such biases could be detected simply by reading transcripts of group deliberations or asking jurors to recount their reasons post-trial, and the JT researchers fail to elaborate how that could be done. The definitions of the various kinds of impermissible reasoning incorporate (through the definition of 'impermissible reasoning') the notion that such reasoning is inherently illogical, implying that it is consciously manifested, but that also is not supported by the cases. On the contrary, judges often advert to the fact that reasoning based on propensity is not objectionable because it is illogical, but because the propensity evidence tends to be given too much weight,[201] a conclusion that is also supported by psychological research.[202] Similarly, the argument for corroborative evidence of any kind is that the accumulation of such evidence tends to strengthen the case for which it is tendered, so accumulation prejudice based on multiple witnesses only differs from permissible reasoning in terms of degree.[203] There is nothing in the Report to explain how this difference in degree was detectable in the JT Study.

The fact that a number of findings suggested irrational use of the evidence undermines the suggestion that the post-trial analysis of deliberations was capable of detecting illogical reasoning. The anomalous divergences between verdicts on the same complaint, between verdicts on joint trials and tendency trials on identical evidence, and between joint trials with and without corroborative evidence, are difficult to explain as a rational outcome. In view of these results, the following finding seems to highlight the insensitivity of the methodology rather than to support the JT researchers' conclusions: 'None of the trials in which tendency evidence was admitted prompted any instances of these three types of impermissible reasoning. None of the juries featured a juror who reasoned illogically about the evidence.'[204]

199 Miles Hewstone, Wolfgang Stroebe and Klaus Jonas, *An Introduction to Social Psychology* (BPS Blackwell, 5th ed, 2012) 94, 102; Robinson, above n 195.

200 *Perry v The Queen* (1982) 150 CLR 580, 594 (Murphy J).

201 *DPP v Boardman* [1975] AC 421, 456 (Lord Cross). See also Law Reform Commission, above n 173, vol 1, 451–60 [795]–[809], cited in Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 42.

202 Robinson, above n 195, 199–204.

203 See Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 47, and references there cited.

204 Ibid 121.

Despite the fact that the JT researchers promote the authenticity of their experimental paradigm, their jury studies lack ecological validity in key respects. The trials were abnormally short and their simulated nature lacked the emotional impact that may well be an essential antecedent or trigger to prejudice. In addition, the participants knew at all times that their jury-room deliberations would be scrutinised ex post facto, which may have inhibited the explicit manifestations of prejudice that the JT Study aimed to detect. The JT researchers have to some extent recognised these shortcomings.[205]

## VII CONCLUSIONS

The JT Study aimed to test various hypotheses underlying the legal perception that joinder of charges of multiple complainants in a single trial risks prejudice to the accused. The hypothetical mechanisms tested were character prejudice, accumulation prejudice, and inter-case conflation of evidence, which are also cited as reasons for caution in admitting evidence of prior conduct.

The joinder effect of concern in practice is the effect arising from a joint trial of all complainants' allegations compared to a single trial with only one complainant. However, the Report frequently confuses this effect with an effect that might be called the joinder effect per se, which is the effect of joining all charges in a single trial, compared to admitting the evidence of the other complainants without joinder of their charges. This effect is not an issue in practice, although it should be observed that if there is a joinder effect per se, it potentially operates in joinder cases in addition to any prejudicial effect of admitting the other complainants' evidence.

Character prejudice is said to arise when extrinsic misconduct by the defendant is used to conclude that the defendant is a person of bad character, and therefore likely to be guilty of the current charges. It therefore operates, if at all, on the admission of evidence of extrinsic conduct, whether by joinder or otherwise. By definition, inter-case conflation of evidence also arises, if at all, out of such admission. The JT researchers consider two forms of accumulation prejudice — prejudice due to accumulation of counts, and prejudice due to accumulation of evidence. The latter is also dependent on the addition of extra evidence, whereas the former is based on multiple counts, which could be regarded as a joinder effect per se.

The JT Study provides no significant support for the theory of inter-case conflation of evidence, but the abbreviated form of the simulated trials raises doubts about whether these findings can be generalised to real life scenarios with much longer and more complex trials and more opportunity for confusion. The same qualification could be made about findings on factual errors and recall.

205 Ibid 268–70.

There was also no support for a theory of prejudice due to accumulation of witnesses.

Despite contrary interpretations by the JT researchers, it is submitted that the JT Study does provide support for some sort of prejudice arising from the joinder effect per se, which might be explained by a theory of accumulation of counts. In trials with identical evidence but a difference in counts, judgements relating to guilt were consistently and significantly more adverse to the accused. When the counts were joined, juries did spread their deliberation time more across the complaints, with the weakest case receiving the most attention and the focal case receiving less than when it was the only case subject to charges.[206] It is possible that the spreading of cognitive effort across the charges increased the salience of the tendency pattern, compared to when the tendency evidence was corroborative only.

In relation to character prejudice, as expected, the addition of tendency evidence did increase the rate of guilty verdicts and other measures reflecting guilt, but the attempts by the JT researchers to rule out such prejudice were, it is submitted, ill-conceived. The quantitative predictions aimed at negativing such prejudice were at least as questionable as the effect they were devised to test, and the analysis of transcripts was incapable of detecting such prejudice for the reasons summarised above. Judicial directions on how to use tendency evidence seemed to have had no systematic effect.

The JT Study also sought to test the impact of relationship evidence — in this case, evidence of alleged grooming behaviours — both with and without special judicial directions. The results suggest that the relationship evidence here did not automatically appeal to juries as a reliable probative adjunct when left to their own cognitive resources. This may be due to the nature of the relationship evidence in the JT Study, which came from the complainant himself and was open to innocent interpretations. The fact that a judicial direction on its use led to a spike in guilty verdicts which was obliterated by the addition of a question trail supports the JT researchers' conclusion that the direction created confusion which the question trail clarified. There was little found to suggest that the relationship evidence logically assisted jury reasoning, or that it was prejudicial, and the supposed rational basis of such evidence deserves some reconsideration.

Another form of judicial direction tested in the JT Study was the question trail. Given the limited types of case in which question trails were used, and the confusing effect of the relationship direction, it is difficult to draw broader conclusions about the use of question trails. When a question trail was used in relationship cases, it merely returned similar outcomes to trials without any judicial directions at all. Its main effect seemed to be to counter the confusion induced by the relationship direction. Question trails also appeared to have no systematic effect in joint trials. However, the same caution could be expressed here as with the effects on memory recall. There is some evidence that question

206 Ibid 186.

trails may help to overcome confusion, and the short duration of the trials in this study may have been inadequate to demonstrate those benefits.

## VIII FUTURE RESEARCH

In this author's view, it goes without saying that scientific and statistical methods can contribute substantially to legal policy debate.[207] The JT researchers had the admirable goal of undertaking an authentic, ecologically valid study of jury reasoning in this complex area. It was a huge project, with over a thousand participants, multiple trial configurations and voluminous analysis and reporting. Nevertheless, their study raises serious questions about the use of simulated trials to test for prejudice of the kind investigated, especially with respect to authenticity. To the extent that prejudice is hypothesised to arise from an emotional response, a design in which the participants know the evidence to be fictional is ill-suited to generating such a response. This may be difficult to cure, because blinding the participants to the fiction may raise ethical issues. Similarly, if the prejudice is said to arise from cognitive errors such as memory lapse, evidentiary conflation or confusion, the shortness of the trials is likely to minimise such effects. In theory, the simulations could be elaborated and lengthened, but apart from issues of costs and logistics, the increased body of evidence would be likely to introduce extraneous variables which might be hard to control.

An insight gained from the JT Study is the critical need for future research to clearly define rational reasoning about tendency and relational evidence, such that it can be distinguished from irrational prejudice in the experimental setting. This is normally the function of 'operationalising' the variables, but the JT researchers' efforts at operationalisation, based on judicial formulations, led only to general hypotheses which were not readily testable by the experimental method employed.[208] Any definition of the supposed prejudice should be informed by psychological as well as judicial theory and must recognise that the prejudice may operate only at the unconscious level.

More generally, the JT Study raises questions about reliance on arbitrary statistical paradigms for legal policy debates of this kind. In the real world, the view that one can wait for a 95 per cent + probability level before accepting a proposition is impractical not only because it sets the bar too high, but also because it ignores the nature of the problem and assumes that the alternative is an acceptable default. In the JT Study, the proposition that joinder has a prejudicial effect was treated as a scientific hypothesis requiring greater than 95 per cent likelihood to be accepted, yet the alternative, that there is no prejudicial effect, required only 5 per cent. This accords with the scientific paradigm that 'no-effect' is treated as the default,

207 Robinson, above n 195; Peter Robinson, 'Graphic and Symbolic Representation of Law: Lessons from Cross-Disciplinary Research' (2009) 16(1) *eLaw Journal: Murdoch University Electronic Journal of Law* 53.

208 Goodman-Delahunty, Cossins, and Martschuk, *Jury Reasoning in Joint and Separate Trials of Institutional Child Sexual Abuse: An Empirical Study*, above n 6, 45–8.

or what the scientists would call the null hypothesis, but it gives the 'no-effect' hypothesis a privileged status inconsistent with the fact that it does not represent the real-world status quo, nor the fact that defendants are entitled to the benefit of any doubt. It is submitted that for legal policy debate, the likelihood threshold for acceptance of any hypothesis must be determined by reference to the nature of the real life problem and the potential effect of false positives or negatives.[209] At least under our current system, a false guilty verdict is regarded as more aversive than failure to convict a guilty offender. If that is to be modified, it should occur transparently through re-calibration of standards, not subversively through blind adherence to arbitrary statistical thresholds. It may be that a risk of prejudice well short of 50 per cent, let alone 95 per cent, would be regarded as unacceptable.

The main novel finding arising from this review was the joinder effect per se. Assuming that this effect can be replicated, its explanation is obscure, and further research on its cause may provide insight into jury reasoning more generally. I have suggested one explanation, based on increased salience, but other explanations are possible.[210] It may be that the failure of the prosecution to lay charges suggests to jurors that the evidence is less strong, or that the uncharged events are less important. Both these theories are consistent with the finding that in tendency trials, juries spend less time on the uncharged events than in joint trials where charges were laid. In fact, perceptions of lack of importance or weakness of evidence may cause the reduced deliberation times and lower salience of the tendency pattern.

Another possibility is that the joinder of charges affords, in the minds of jurors, a kind of compromise verdict, in which they respond to moderate uncertainty by convicting on some charges but not others. In the tendency trials where only one case is the subject of charges, no such compromise is possible, so jurors may simply acquit. A compromise theory could also explain differences in verdicts between major and minor charges with the same complainant.

The effect of the judicial directions and question trails in the relationship trials might also be explained by relative salience. The relationship direction made relationship reasoning salient, whereas the question trail brought jurors' minds back to the basic questions at the core of their decision making process. It may be possible to design studies which test the effect of salience on these reasoning processes. In fact, salience might be used as a control for activating or deactivating particular evidence or modes of reasoning to test for their effects.

As a final suggestion, it is to be hoped that, like the transcripts of the trials, the complete dataset used in the main study and pilot, as well as the deliberation transcripts, will be made publicly available. The JT Study not only contributed to the body of knowledge by providing the findings in the Report, but also by generating a rich dataset. Analysis of the published data suggests many more analytical possibilities not covered either in the Report or this article, and matters that seem of secondary importance today may become of interest in future years

209 In null hypothesis testing, these are called Type I and Type II errors.

210 For these suggestions, I am indebted to a reviewer of the *Monash University Law Review*.

or decades. As the preservation of datasets in an increasingly electronic world becomes more routine, their contribution to the body of knowledge will hopefully be recognised by public agencies through an increased willingness to fund such research, not only for the purpose of current policy debate, but also for a future generation of researchers striving not to re-invent the wheel.