

## ARTICLE OPEN

(Nearly) optimal  $P$  values for all Bell inequalitiesDavid Elkouss<sup>1</sup> and Stephanie Wehner<sup>1</sup>

A key objective in conducting a Bell test is to quantify the statistical evidence against a local-hidden variable model (LHVM) given that we can collect only a finite number of trials in any experiment. The notion of statistical evidence is thereby formulated in the framework of hypothesis testing, where the null hypothesis is that the experiment can be described by an LHVM. The statistical confidence with which the null hypothesis of an LHVM is rejected is quantified by the so-called  $P$  value, where a smaller  $P$  value implies higher confidence. Establishing good statistical evidence is especially challenging if the number of trials is small, or the Bell violation very low. Here, we derive the optimal  $P$  value for a large class of Bell inequalities. What is more, we obtain very sharp upper bounds on the  $P$  value for all Bell inequalities. These values are easily computed from the experimental data, and are valid even if we allow arbitrary memory in the devices. Our analysis is able to deal with imperfect random number generators, and event-ready schemes, even if such a scheme can create different kinds of entangled states. Finally, we review requirements for sound data collection, and a method for combining  $P$  values of independent experiments. The methods discussed here are not specific to Bell inequalities. For instance, they can also be applied to the study of certified randomness or to tests of noncontextuality.

npj Quantum Information (2016) 2, 16026; doi:10.1038/npjqi.2016.26; published online 18 October 2016

## INTRODUCTION

Local-hidden variable models (LHVM) predict concrete limitations on the statistics that can be observed in a Bell experiment.<sup>1</sup> These are typically phrased in terms of probabilities or expectation values. However, in any experiment we can only observe a finite number of trials, and not probabilities. We thus need to quantify the statistical evidence against an LHVM given a finite number of trials.

The traditional way to analyse statistics in Bell experiments is to compute the number of standard deviation that separate the observed data from the best LHVM. However, it is now known that this method has flaws<sup>2–5</sup> (see ref. 4 for a detailed discussion). In particular, we would have to assume Gaussian statistics and independence between subsequent attempts, allowing for the memory loophole.<sup>2,3</sup> Fortunately, it is possible to rigorously analyse the statistical confidence even when allowing for memory as was first done by Gill.<sup>6</sup> This is the approach that we follow here.

Instead of bounding the standard deviation, the intuitive idea behind the rigorous analysis is to bound the probability of observing the experimental data if nature was indeed governed by an LHVM. In the language of hypothesis testing, this is known as the  $P$  value, where the null hypothesis is that the experiment can be modelled as an LHVM (see e.g., ref. 7). Informally, we thus have

$$P\text{-value} = \max_{\text{LHVM}} \Pr[\text{data at least as extreme as observed} \\ \mid \text{experiment is governed by LHVM}]. \quad (1)$$

A small  $P$  value can be interpreted as strong evidence against the null hypothesis. Hence, in the case of a Bell experiment, a small  $P$  value can be regarded as strong evidence against the hypothesis that the experiment was governed by an arbitrary LHVM.

There is extensive literature regarding the methods for evaluating the  $P$  value in Bell experiments<sup>2–16</sup> and discussions regarding the analysis of concrete experiments and loopholes.<sup>17–32</sup> Previous approaches to obtain such  $P$  values

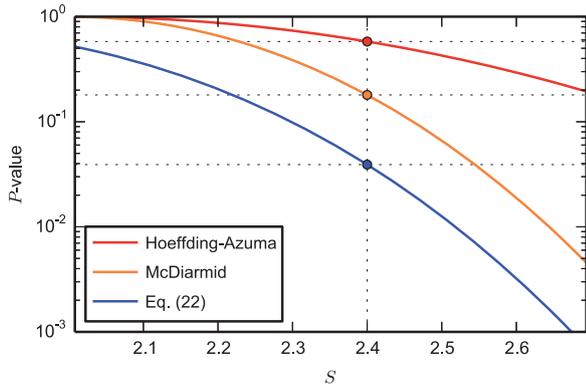
known from the literature can be roughly divided into two categories. In the first approach, we select a suitable Bell inequality based on the expected experimental statistics or test data collected ahead of time. After a Bell inequality is fixed, one can model the process as a (super-)martingale to which standard concentration inequalities<sup>2,6,9–11,13,15</sup> can be applied. Although this allows one to obtain bounds for all Bell inequalities relatively easily, the resulting upper bounds on the  $P$  values are generally very loose. Crucially, this means that a much larger amount of trials would need to be collected than is actually necessary to obtain good statistical confidence. Figure 1 illustrates the significance of using bounds employed in previous works compared with the bound used here. When making a statement about all Bell inequalities below, we will also take a martingale approach using the much sharper concentration offered by the Bentkus' inequality.<sup>33</sup> For some simple inequalities like Clauser–Horne–Shimony–Holt (CHSH)<sup>34</sup> and Clauser–Horne (CH),<sup>35</sup> tight bounds on the  $P$  value have been obtained when the measurement settings in the experiment are chosen uniformly, and no event-ready scheme is employed.<sup>3,5,16</sup> Such a bound was first informally derived in,<sup>3</sup> and later rigorously developed by Bierhorst,<sup>5</sup> whose approach for CHSH closely inspires our analysis of Bell inequalities that correspond to win/lose games below.

The second approach that has been pursued is to combine the search for a good Bell inequality with a numerical method adapting to the data.<sup>4,12,14</sup> This method is asymptotically optimal in the limit of many experimental trials. Although conceptually beautiful, this numerical method can need a rather significant amount of trials to out-perform even the somewhat loose bounds given by standard martingale concentration inequalities, and can hence only be used in regimes where the amount of trials collected in the experiment is indeed large.

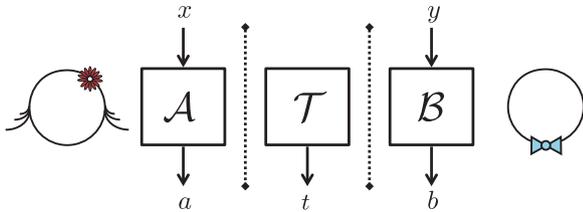
<sup>1</sup>QuTech, Delft University of Technology, Delft, The Netherlands.

Correspondence: S Wehner (s.d.c.wehner@tudelft.nl)

Received 3 November 2015; revised 14 March 2016; accepted 2 April 2016



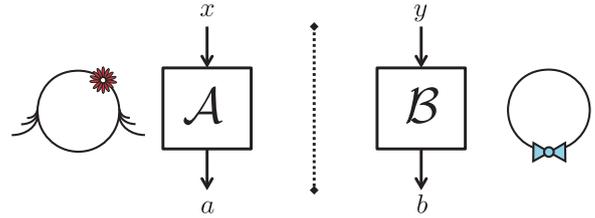
**Figure 1.** Comparison of  $P$  value bounds for the CHSH inequality for values used in the first loophole-free Bell test.<sup>36</sup> The three curves show bounds on the  $P$  value for a fixed number of trials  $n=245$  and random number generators bias  $\tau=1.08 \cdot 10^{-5}$ . (refs 53,54) The  $P$  value is computed as a function of the violation  $S$ , which is defined as  $S=8(c/n-1/2)$ , where  $c$  is the number of wins in the CHSH game. From top to bottom, the curves show the bound on the  $P$  value computed with Azuma–Hoeffding used in, (ref. 11) McDiarmid’s inequality<sup>40</sup> given in ref. 14 and the upper bound from (20) (with  $\beta_{\text{win}}=3/4+\tau-\tau^2$  as shown in Lemma 1 in the Supplemental Material Section II). In the Delft experiment<sup>36</sup> a number  $c=196$  of wins were observed, giving  $S=2.4$  and (20) yields  $P$  value  $\approx 0.039$ . The dots indicate the  $P$  values predicted by the other bounds. To obtain the same  $P$  value with McDiarmid’s inequality and Azuma–Hoeffding<sup>11</sup> the required violations would be  $S=2.54$  and  $S=2.98$  (beyond QM), respectively.



**Figure 2.** A Bell test using an event-ready scheme as proposed by Bell.<sup>1,59</sup> In an event-ready scheme, there is an additional site that we call the ‘heralding station’ that is space-like separated from Alice and Bob at the time they receive their inputs (see Figure 3). This heralding station can be under full control of the local-hidden variable model. It takes no input but produces a tag  $t$  as output. In the simplest case,  $t$  is just a single bit where  $t=1$  corresponds to ‘yes’ and  $t=0$  to ‘no’. If yes, then we check the winning condition for Alice and Bob as in Figure 3. If no, then no record is made (i.e., the null game is played). In physical implementations such as<sup>36</sup> this tag indicates whether an attempt to produce entanglement was successful. More complicated scenarios are possible, in which the tag  $t$  takes on more than two-values. Depending on  $t$ , a particular game is played – i.e., scores are computed as dictated by the game labelled by  $t$ . In physical implementations this is interesting when two different entangled states can be created in the event-ready scheme, and each state is best for a particular game. An example is given by CHSH, where different Bell states are created and we play two different CHSH games with  $x \cdot y = a \oplus b$  or  $x \cdot y = a \oplus b \oplus 1$ . Using both states can improve the time scales at which statistical confidence can be obtained.

## RESULTS

Here we present a method for analysing the  $P$  value for Bell experiments that is optimal for large classes of Bell inequalities. This method also applies to event-ready schemes as used in,<sup>36</sup> and can also deal with more complicated forms of event-ready procedures (heralding) in which different states are created in each trial (see Figure 2). In particular, it applies to situations in which we apply a different Bell inequality at each trial depending on which state is generated. Furthermore, we show how to bound



**Figure 3.** A Bell test involving two space-like separated sites, labelled Alice and Bob. Alice and Bob receive two randomly chosen inputs  $x$  and  $y$ , and produce outputs  $a$  and  $b$ . We indicate that Alice and Bob are space-like separated via the dotted line. When testing the CHSH inequality, for example, the inputs and outputs can be taken to be single bits  $x, y, a, b \in \{0, 1\}$ . Viewing CHSH as a non-local game, the winning condition is that  $x \cdot y = a \oplus b$  (we use the shorthand  $a \oplus b$  to denote  $a+b \bmod 2$ ). This means that in one trial of the experiment, we check whether  $x \cdot y = a \oplus b$  and if yes we increment the number  $c$  of wins by 1. For all Bell inequalities that are win/lose games (see ‘ $P$  values for win/lose games’), we analogously count the number of wins. General Bell inequalities (see ‘General games’) can also be cast as a game in which we do not just decide on whether Alice and Bob win or lose, but instead assign a score to each correct answer. In the experiment, we then compute the total score from the inputs and outputs observed. Our analysis is analogous for Bell inequalities involving more than two sites.

the  $P$  value of all Bell experiments using Bentkus’ inequality, which is optimal up to a small constant.

Before we can state the concept of a  $P$  value more precisely, let us briefly recall the concept of a Bell inequality (see e.g., (ref. 37) for an in-depth introduction). For simplicity, we thereby restrict ourselves to Bell inequalities involving two sites (Alice and Bob), but all our arguments hold analogously for an arbitrary number of sites. As illustrated in Figure 3, in a Bell experiment we choose inputs  $x$  and  $y$  to Alice and Bob, and can record outputs  $a$  and  $b$ . (Note that we here use the more common notation of  $a$  and  $b$  being outputs, and  $x$  and  $y$  being inputs. The roles are reversed in ref. 36) If the experiment was governed by an LHVM, we could write the probabilities of obtaining outputs  $a$  and  $b$  given inputs  $x$  and  $y$  as

$$p(a, b|x, y) = \int d\mu(h) p(a|x, h) p(b|y, h), \quad (2)$$

where  $d\mu$  is an arbitrary measure over hidden variables  $h$ , which also include any prior history of the experiment. The locality of the model is captured by the fact that  $p(a, b|x, y, h) = p(a|x, h) p(b|y, h)$  if Alice and Bob are indeed space-like separated. Throughout, we refer to the Supplemental Material for a formally precise notation, definitions and derivation. A Bell inequality then states that for any LHVM

$$\beta_{\min} \leq \sum_{x,y,a,b} s_{ab}^{xy} p(a, b|x, y) \leq \beta_{\max}, \quad (3)$$

for some numbers  $s_{ab}^{xy}$ . Evidently, in an experiment we never have access to actual probabilities  $p(a, b|x, y)$ . Nevertheless, Bell inequalities turn out to be very useful to establish bounds on the  $P$  value above.

Let us now rephrase this inequality in a way that will make our approach more intuitive later on. In an experiment we choose settings with some probability  $p(x, y)$ ; hence, it will be convenient to define

$$s_{ab|xy} = s_{ab}^{xy} / p(x, y). \quad (4)$$

For the moment, let us assume we have perfect random number generators, and that we choose the settings  $x$  and  $y$  uniformly such that  $p(x, y) = p(x)p(y)$  where  $p(x) = 1/N_x$  and  $p(y) = 1/N_y$ . The Bell inequality then reads

$$\beta_{\min} \leq \frac{1}{N_x N_y} \sum_{x,y,a,b} s_{ab|xy} p(a, b|x, y) \leq \beta_{\max}. \quad (5)$$

The reason why this notation is convenient is because we can now think of  $s_{ab|xy}$  as a score that Alice and Bob obtain when giving answers  $a$  and  $b$  for questions  $x$  and  $y$ . We thus adopt a modern formulation of Bell inequalities in terms of games.<sup>37</sup> The statement that an LHV governs the experiment then means that Alice and Bob can only use a local-hidden variable strategy to achieve a high score in the game. Using this formulation it is clear that the term in (5) is just the average score that Alice and Bob can hope to achieve in the next trial. As the Bell inequality holds for any local-hidden variables, including the history, it is clear that playing the game  $n$  times in succession – i.e., performing  $n$  trials of the experiment – corresponds to a classic example of martingale sequence (Supplementary Material).

To analyse the experimental data we then proceed as follows: In trial  $j$ , we compute the score  $s_{a,b_j|x_j,y_j}$  that Alice and Bob obtain for the inputs  $x$  and  $y$  and outputs  $a$  and  $b$  we observed in that trial. By adding all these numbers we compute the total score  $c = \sum_{j=1}^n s_{a,b_j|x_j,y_j}$  after performing  $n$  trials. The  $P$  value then corresponds to

$$P\text{-value} \leq \max_{\text{LHVM}} \Pr[\text{Alice and Bob score } C \geq c | \text{LHVM}]. \quad (6)$$

That is, the probability that Alice and Bob would obtain a score  $C$  that is at least as large ( $C \geq c$ ) as the score  $c$  actually observed in our experiment.

Note that the choice for the score function, corresponding to a particular Bell inequality, is not unique. The only restriction, in order to define a  $P$  value, is that the score needs to be a valid test statistic. A test statistic is a function that assigns a real value to each possible experimental outcome. Then, the  $P$  value is the probability, under the null hypothesis, that the value of the test statistic is equal to or larger than the value obtained from the observed data. There are many possible score functions that verify this restriction, though we would argue that the one used here is particularly natural.

#### $P$ values for win/lose games

We first obtain optimal  $P$  values for a certain class of Bell inequalities, also known as *non-local games*. In particular, this includes the Bell inequalities phrased in terms of correlation functions such as the famous CHSH inequality.<sup>34</sup> What sets these inequalities apart is that the scores  $s_{ab|xy}$  can take on only two values, which we associate with winning or losing the game.

*Winning probability.* To illustrate how Bell inequalities correspond to games, let us consider the CHSH correlation function

$$\langle A_0 B_0 \rangle + \langle A_1 B_0 \rangle + \langle A_0 B_1 \rangle - \langle A_1 B_1 \rangle, \quad (7)$$

where  $A_x$  and  $B_y$  correspond to the observables measured by Alice and Bob, respectively (see Figure 3). Note that we can write one of the correlators as

$$\langle A_x B_y \rangle = \sum_a p(a, b = a | x, y) - \sum_a p(a, b = a \oplus 1 | x, y). \quad (8)$$

In terms of the score function, this means that  $s_{a,b|x,y} = (-1)^{a+b+xy}$ . Note that in any game in which  $s_{a,b|x,y}$  can only take on these two values we can think of the probability that Alice and Bob win for a particular choice of measurement settings  $x$  and  $y$  as

$$p(\text{win} | x, y) = \sum_{\substack{a, b \\ s_{a,b|x,y} = 1}} p(a, b | x, y), \quad (9)$$

$$p(\text{lose} | x, y) = \sum_{\substack{a, b \\ s_{a,b|x,y} = -1}} p(a, b | x, y) \quad (10)$$

$$= 1 - p(\text{win} | x, y). \quad (11)$$

To draw full analogy with the usual representation of non-local games (see e.g. see ref. 37) we can normalise any score function for which  $s_{a,b|x,y} \in \{\pm 1\}$  to be 0 and 1 by defining  $\hat{s}_{a,b|x,y} = (s_{a,b|x,y} + 1)/2$ .

We have

$$p(\text{win}) = \sum_{x,y} p(x, y) p(\text{win} | x, y), \quad (12)$$

which is precisely the probability that Alice and Bob win the non-local game.<sup>37</sup> In this language, a Bell inequality now takes on the form

$$p(\text{win}) \leq \beta_{\text{win}} \quad (13)$$

where  $\beta_{\text{win}}$  denotes the optimal winning probability that can be achieved using an LHV. Note that, if necessary,  $\beta_{\text{win}}$  can be obtained by normalising the given values  $\beta_{\text{min}}, \beta_{\text{max}}$  appropriately.

*Analysing data.* The following steps need to be taken to obtain a  $P$  value for an experiment based on a non-local game, where for simplicity we first consider schemes that are not event-ready. We refer to Section I of the Supplementary Material for details and derivation.

First, we determine a bound on the bias of the random number generator. We will never be able to generate settings  $x$  and  $y$  exactly according to the specific distributions  $p(x)$  and  $p(y)$ . Instead we will generate the settings according to some other distributions,  $\tilde{p}(x)$  and  $\tilde{p}(y)$ . We are interested in the numbers  $\tau_A$  and  $\tau_B$  such that

$$|p(x) - \tilde{p}(x)| \leq \tau_A, \quad (14)$$

$$|p(y) - \tilde{p}(y)| \leq \tau_B. \quad (15)$$

It is clear that for any physical device, these are estimates ideally supported by a theoretical device model with clearly specified assumptions.

Second, we need to obtain a bound on the winning probability conditioned on the history of the experiment. This bound should use such imperfect random number generators (RNGs) and be valid for all LHVMs

$$\tilde{p}(\text{win} | \text{History}) = \sum_{x,y} \tilde{p}(x, y) p(\text{win} | x, y) \leq \tilde{\beta}_{\text{win}}. \quad (16)$$

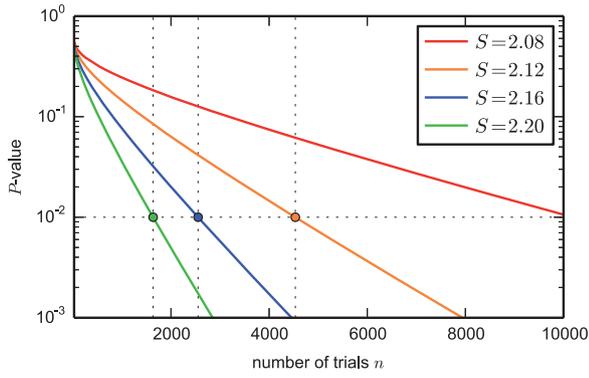
Such a bound can be obtained analytically for many inequalities, including CHSH (see Section II of the Supplementary Material). In general, a bound on  $\tilde{\beta}_{\text{win}}$  can be computed numerically using a linear program (LP), when re-normalising the score functions  $\hat{s}_{a,b|x,y} \in \{0, 1\}$  as above. We remark that this LP has size that is exponential in the number of inputs and outputs, but can nevertheless be solved numerically when these are small enough, which is typically the case in all experimental Bell tests. It is known that it is NP-hard to compute the winning probability for arbitrary non-local games.<sup>38</sup>

Third, in each of the  $n$  experimental trials, we generate inputs  $x$  and  $y$  and record outputs  $a$  and  $b$ . In the end, we count the number of trials,  $c$ , in which Alice and Bob won the game – i.e., the number of times  $s_{ab|xy} = 1$ .

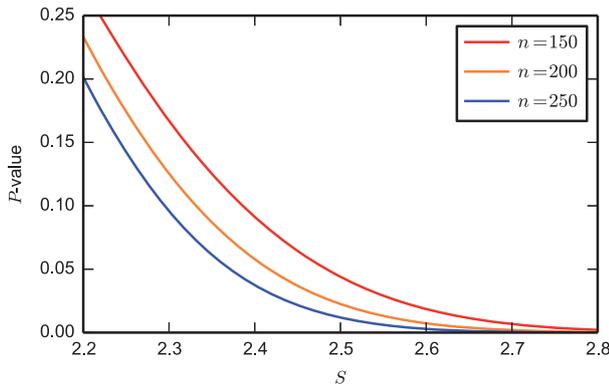
Finally, we compute the  $P$  value. The interpretation of the  $P$  value is the probability that Alice and Bob win at least  $c$  times, maximised over any LHV strategy.

$$P\text{-value} = \max_{\text{LHVM}} \Pr[\text{Alice and Bob win at least } c \text{ times} | \text{LHVM}]. \quad (17)$$

As we prove in Lemma 3 of the Supplementary Material, for all LHVMs including arbitrary memory effects,



**Figure 4.**  $P$  values for the CHSH inequality with imperfect random number generators (the bias is  $\tau = 1.08 \cdot 10^{-5}$ ) in regimes where the violation is very low, but the number of trials is large. The  $P$  values are computed with (20). The curves show the  $P$  value as a function of the number of trials for fixed violation values:  $S = 2.08$ ,  $S = 2.12$ ,  $S = 2.16$  and  $2.20$ . The dashed horizontal line is set at  $P$  value = 0.01. This line is crossed at  $n = 10195$ ,  $n = 4534$ ,  $n = 2552$  and  $n = 1635$  trials, respectively.



**Figure 5.**  $P$  values for the Mermin's inequality<sup>39</sup> with perfect random number generators. Mermin's inequality is a tripartite inequality in which each party has two inputs and two possible outputs. It is an example of a non-bipartite inequality that has already been violated in the laboratory.<sup>55–57</sup> The three parties Alice, Bob and Charlie receive three random chosen inputs  $x, y$  and  $z$  with the promise that the parity of the inputs is even—that is, that the inputs are limited to  $(0, 1, 1)$ ,  $(1, 0, 1)$ ,  $(1, 1, 0)$ ,  $(0, 0, 0)$ , and produce outputs  $a, b$  and  $c$ , which can also be taken to be bits—that is,  $x, y, z, a, b, c \in \{0, 1\}$ . The winning condition for Mermin's inequality is that  $a \oplus b \oplus c = x \vee y \vee z$ . That is the game is won if the xor of the outputs equals 0 when  $(x, y, z) = (0, 0, 0)$  and if the xor of the outputs equals 1 in the remaining cases. Hence, we get  $s_{abc|xyz} = a \oplus b \oplus c + 1$  when  $(x, y, z) = (0, 0, 0)$  and  $s_{abc|xyz} = a \oplus b \oplus c$  when  $(x, y, z) \neq (0, 0, 0)$ . The winning probability for this game is  $p(\text{win}) = 3/4$ ,<sup>58</sup> but note that in contrast with CHSH if Alice, Bob and Charlie share entanglement they can win with probability one. The curves show the  $P$  value as a function of  $S = 8(c/n - 1/2)$  for a fixed number of trials  $n$  ( $c$  is the number of wins). The three curves show from top to bottom the  $P$  value for  $n = 150$ ,  $n = 200$  and  $n = 250$ . The  $P$  values are computed with the binomial upper bound (18).

$$P\text{-value} \leq \sum_{i=c}^n \binom{n}{i} (\tilde{\beta}_{\text{win}})^i (1 - \tilde{\beta}_{\text{win}})^{n-i}. \quad (18)$$

This bound is a generalisation of (ref. 3) and (ref. 5) that already had given a binomial upper bound for one particular win/lose game, the CHSH game, when the RNGs are perfect, and no event-ready scheme is used.

We emphasise that this bound is tight, whenever (16) is tight. That is, there exists an LHM that produces at least  $c$  wins with this probability, and this LHM does not use any memory. Although a theoretical analysis is of course necessary to prove (18), it follows that the memory loophole<sup>3</sup> can only be exploited for general Bell inequalities, where it indeed turns out to be significant. Figure 4 and Figure 5 illustrate this bound for the CHSH and Mermin's inequality.<sup>39</sup>

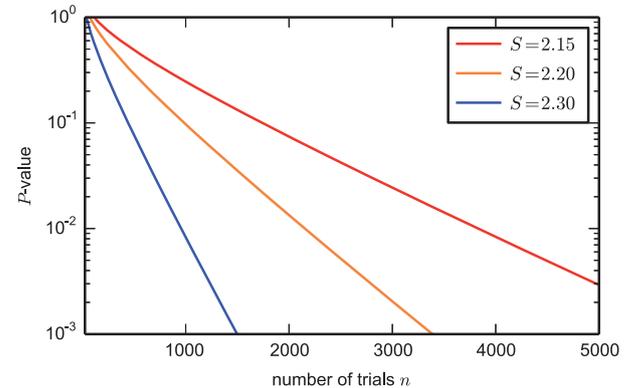
**Event-ready schemes.** To illustrate the analysis of event-ready schemes, let us here focus on the usual case where the tag (see Figure 2) can be either  $t=0$  (null game, no entanglement was made) or  $t=1$  (one game, one specific state was made). We will use the term attempt to refer to an attempt to create entanglement (outcome  $t=0$  or  $t=1$ ) and reserve the word trial for those in which  $t=1$ . In Section II of the Supplementary Material, we will discuss more complex versions of event-ready schemes in which different entangled states can be created, and we employ a different game for each state.

While it is important that the random numbers are chosen independently of the tag  $t$ , we otherwise allow the LHM arbitrary control over the statistics of heralding station. In particular, this means that the LHM may use more (or fewer) attempts to realise  $c$  wins on  $n$  trials than we actually observed during the experiment.

Specifically,

$$P\text{-value} = \max_{\text{LHM}} \sum_{t^m \in \{0,1\}^m | |t^m|=n} \Pr[t^m | \text{LHM}] \Pr[C \geq c \text{ wins} | \text{LHM}, t^m], \quad (19)$$

where  $t^m = t_1, \dots, t_m$ ,  $|t^m|$  denotes the number of ones in  $t^m$ , and the maximisation over LHM includes an optimisation over an arbitrarily large number of attempts,  $m$ , and heralding statistics. As we will formally show in the Lemma 3 of the Supplementary Material,



**Figure 6.**  $P$  values for CGLMP's inequality<sup>41</sup> with perfect random number generators. CGLMP is a sequence of bipartite inequalities in which each party has two inputs and  $d \geq 2$  possible outputs. This is an example of a general game within experimental reach.<sup>60</sup> The inequality is  $\sum_{k=0}^{\lfloor d/2-1 \rfloor} \sum_{a,x,y} (1 - \frac{2k}{d-1}) (p(a, a+k+xy|x,y) - p(a, a-k-1+xy|x,y)) \leq 2$ . Let  $k \in 0 \dots \lfloor d/2-1 \rfloor$ , we can extract from the inequality the score functions  $s_{ab|xy} = 4(1 - \frac{2k}{d-1})$  if  $b = a+k+xy$ ,  $s_{ab|xy} = -4(1 - \frac{2k}{d-1})$  if  $b = a-k-1+xy$ , and  $s_{ab|xy} = 0$  in the remaining cases. The three curves show the  $P$  value as a function of the number of attempts for a fixed average score  $S = \frac{1}{n} \sum_{j=1}^n s_{a_j b_j | x_j y_j}$ . From top to bottom the curves show the  $P$  value for  $S = 2.15$ ,  $S = 2.20$  and  $S = 2.30$ . The  $P$  values are computed via Bentkus' inequality (25).

$$P\text{-value} = \sum_{i=c}^n \binom{n}{i} (\tilde{\beta}_{\text{win}})^i (1 - \tilde{\beta}_{\text{win}})^{n-i}. \quad (20)$$

That is, we can formally ignore the non-successful attempts. The  $P$  value only depends on the trials.

### General games

Let us now move on to considering general games—that is, games in which the score functions  $s_{ab|xy}$  take on more than two possible values (see Figure 6). As before, we first need to consider the bias. Our bound will depend on the values of

$$s_{\text{max}} = \max_{a,b,x,y} s_{ab|xy}, \quad (21)$$

$$s_{\text{min}} = \min_{a,b,x,y} s_{ab|xy}. \quad (22)$$

Recall that as  $s_{ab|xy} = s_{ab}^{xy}/p(x,y)$  the distribution  $p(x,y)$  and hence also the bias influence  $s_{\text{max}}$  and  $s_{\text{min}}$ . Second, we again compute the total score

$$c = \sum_{j=1}^n s_{a_j b_j | x_j y_j}, \quad (23)$$

where  $x_j$  and  $y_j$ , and  $b_j$  and  $a_j$  are the inputs and outputs used during trial  $j$ , respectively. We then have that

$$P\text{-value} = \max_{\text{LHVM}} \Pr[C \geq c | \text{LHVM}], \quad (24)$$

where  $C$  is the random variable corresponding to obtaining a particular score using the LHVM strategy. Using the Bentkus' inequality, we prove in Section III of the Supplementary Material that

$$P\text{-value} \leq e \left( \left( \sum_{i=\lfloor \delta \rfloor}^n \binom{n}{i} (\hat{\gamma})^i (1 - \hat{\gamma})^{n-i} \right)^{1 - \delta + \lfloor \delta \rfloor} \left( \sum_{i=\lceil \delta \rceil}^n \binom{n}{i} (\hat{\gamma})^i (1 - \hat{\gamma})^{n-i} \right)^{\delta - \lceil \delta \rceil} \right) \quad (25)$$

where

$$\delta = \sum_{i=1}^n \frac{c_i - s_{\text{min}}}{s_{\text{max}} - s_{\text{min}}}, \quad (26)$$

$$\hat{\gamma} = \frac{\beta_{\text{max}} - s_{\text{min}}}{s_{\text{max}} - s_{\text{min}}}. \quad (27)$$

Whenever the Bell inequality is normalised such that  $s_{\text{min}} = 0$  and  $s_{\text{max}} = 1$  this becomes

$$P\text{-value} \leq e \left( \left( \sum_{i=\lfloor c \rfloor}^n \binom{n}{i} (\beta_{\text{max}})^i (1 - \beta_{\text{max}})^{n-i} \right)^{1 - c + \lfloor c \rfloor} \left( \sum_{i=\lceil c \rceil}^n \binom{n}{i} (\beta_{\text{max}})^i (1 - \beta_{\text{max}})^{n-i} \right)^{c - \lceil c \rceil} \right) \quad (28)$$

where  $\lfloor c \rfloor$  and  $\lceil c \rceil$  stand, respectively, for the greatest integer smaller than  $c$  and the smallest integer larger than  $c$ .

If we treat a win/lose game as a general game we can also upper bound the  $P$  value by (28). However, if we compare this formula with (18), we see that we have lost a factor of  $e$ . We have obtained a simple formula that can address general games but it is not tight. It remains unknown whether or not  $e$  is the optimal prefactor, but it is known that for general games it cannot be smaller than  $2$ .<sup>33</sup>

In some cases it is possible to transform a general game into a win/lose game by postselecting the trials that take the maximum

and minimum value.<sup>2,16</sup> In that situation, it would be possible to apply the tight bounds for win/lose games. Techniques sometimes referred to as 'speeding up time'<sup>2,28</sup> can analogously be used in conjunction with this refined bound.

The idea behind this bound is to model an experiment as a bounded difference supermartingale, where we note that a Bell inequality is nothing else than the expectation of the score random variable  $C_j$  in trial  $j$  conditioned on the history leading up to that trial. That is,

$$\beta_{\text{min}} \leq \mathbb{E}[C_j | \text{History}] \leq \beta_{\text{max}}, \quad (29)$$

where the expectation is taken over all inputs  $x$  and  $y$  and outputs  $a$  and  $b$ . A (super)martingale is a concept known from probability theory. A sequence  $M_1, M_2, \dots$  of random variables is known as a supermartingale, if the expectation value of the difference  $M_n - M_{n-1}$  conditioned on the history is always non-positive. Choosing  $M_j$  to be a weighted sum of the differences  $\sum_{j=1}^n C_j - \beta_{\text{max}}$  one can easily obtain such a martingale. The key aspect of a martingale is that, even though the subsequent variables are not independent of each other, we observe a concentration akin to the law of large numbers for processes that are independent of each other. The prime example is tossing a coin  $n$  times. Indeed, thinking of 'heads' as 'win' and 'tails' as 'lose', we can easily evaluate the probability that we get 'win' more than  $k$  times. When a process is a martingale a similar argument holds, even if the coin can take many values and depend on the history.

Several other martingale bounds have been used in the past. We have chosen as examples McDiarmid's inequality<sup>40</sup> as given in ref. 14

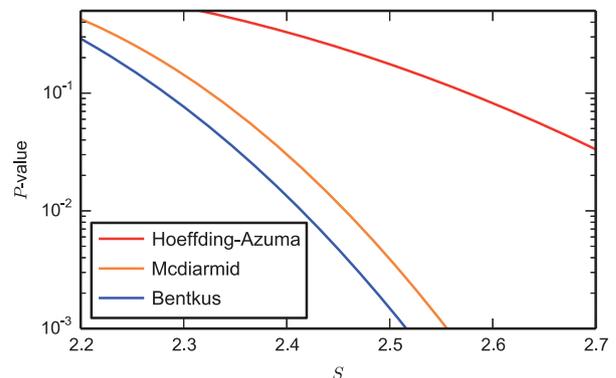
$$P\text{-value} \leq \left( \left( \frac{s_{\text{max}} - \beta_{\text{max}}}{s_{\text{max}} - c/n} \right)^{\frac{s_{\text{max}} - c/n}{s_{\text{max}} - s_{\text{min}}}} \left( \frac{\beta_{\text{max}} - s_{\text{min}}}{c/n - s_{\text{min}}} \right)^{\frac{c/n - s_{\text{min}}}{s_{\text{max}} - s_{\text{min}}}} \right)^n \quad (30)$$

and Azuma–Hoeffding as used in ref. 11 for CHSH

$$P\text{-value} \leq \exp \left( -n \frac{(c/n - \beta_{\text{max}})^2}{2d^2} \right) \quad (31)$$

where  $d = \max \{ |\beta_{\text{max}} - s_{\text{min}}|, |\beta_{\text{max}} - s_{\text{max}}| \}$ .

We provide an example of the application of these three bounds for the Collins–Gisin–Linden–Massar–Popescu inequality<sup>41</sup> in Figure 7.



**Figure 7.**  $P$  values for Collins–Gisin–Linden–Massar–Popescu's (CGLMP) inequality<sup>41</sup> with perfect random number generators. From top to bottom the curves show the  $P$  value for  $n = 500$  trials computed with Azuma–Hoeffding, McDiarmid and Bentkus' inequalities. The inequalities are given by (31), (30) and (25).

## DISCUSSION

Before conducting the experiment

To ensure sound data collection, there are several important considerations to make before the experiment takes place. These are standard in statistical testing, and in essence say that the rules on how the statistical analysis is performed are decided independent of the data. This can be achieved by establishing those rules before the data collection starts. First, we choose a Bell inequality. Not all Bell inequalities lead to the same statistical confidence. In 'Selecting a Bell inequality' we discuss methods for obtaining a good one. Although there may be future analyses that allow a partial optimisation over Bell inequalities using the actual experimental data, we emphasise that the procedure above assumes that a fixed inequality has been chosen ahead of time. Second, there are two ways to deal with imperfect random number generators, and a choice should be made as discussed in 'How to deal with imperfect RNGs'. Third, we assume that the number of trials to be collected is independent of the data. This means that we do not decide to take another few trials if the  $P$  value is not yet low enough for our liking, a practice also known as  $P$  value fishing in statistics. There are ways to augment the analysis<sup>42</sup> to safely collect more data in some specific instances, but this brings many subtleties. A number of trials,  $n$ , can be determined from the expected violation given prior device characterisation, aiming for a particular  $P$  value.

How to deal with imperfect RNGs

From the discussions above, it becomes clear that there are two ways to deal with imperfect RNGs. The first is of interest in win/lose games. If there is a bias  $\tau$ , then the winning probability (12) simply increases. This means that when we perform an experiment based on a win/lose game in which we use an imperfect RNG, the game remains win/lose and the bound in (18) still applies. As this is a simple binomial distribution, without any additional factor  $e$  this is desirable if the bias is small.

However, we saw from the analysis of general games that there is a second way. When considering a general Bell inequality (3), we make no statements about the probabilities of choosing settings  $p(x, y) = p(x)p(y)$ . Starting from a scoring function  $s_{ab}^{xy}$  we can define  $s_{ab|xy} = s_{ab}^{xy}/p(x, y)$  to introduce an explicit dependence on the input distribution  $p(x, y)$  of our choosing. Using RNGs with a bias then merely affects  $p(x, y)$  and thus the maxima and minima of the scoring functions  $s_{ab|xy}$  that enter into the bound given in (25). It is crucial to note that when defining  $s_{ab|xy}$  as above, a win/lose game can now turn into a general game. That is, we will no longer have that the scoring functions  $s_{ab|xy}$  take on only two values. This means that we have to use the general bound (25) carrying the additional factor  $e$ , as opposed to (18).

How we deal with imperfect RNGs thus depends: if we start with a win/lose game, and if the bias is small, it is typically advantageous to preserve the win/lose property of the game and derive a new winning probability as a function of the bias. If, however, the bias is very large, it can be advantageous to sacrifice the win/lose property and adopt the analysis for general games. If the game was not win/lose to begin with, we always adopt the second method.

Selecting a Bell inequality

One of the main objectives of a Bell experiment is to quantify the evidence against an LHVM; hence, ideally one would like to choose a game that would yield the lowest  $P$  value for a fixed number of trials. The optimisation of games with this objective is a non-trivial task. A reasonable alternative that one can use as heuristic is to maximise the gap between the expected score achievable in the experiment and the expected score that an LHVM can attain. In other words, we are looking for a Bell

inequality for which the violation we can observe is as large as possible. To find such an inequality, standard linear programming methods can be used (see e.g., ref. 37).

To apply them we assume that a reasonably good guess is available as to what the probabilities  $p(a, b|x, y)$  are in the experiment. Such a guess can be made by either analysing data collected prior to the Bell experiment and approximating probabilities by relative frequencies or by having sufficient confidence in the theoretical model that describes the experiment and calculating the probabilities from this model.

Suppose that in the estimation process we find some estimates of  $p(a, b|x, y)$ . If such probabilities could be realised by an LHVM, we could write them as a mixture of deterministic local strategies. To make this precise, let  $\lambda = (a_1, \dots, a_{|X|}, b_1, \dots, b_{|Y|})$  denote a deterministic strategy in which Alice and Bob give outputs  $a_x$  and  $b_y$  for inputs  $x = \{1, \dots, |X|\}$  and  $y \in \{1, \dots, |Y|\}$ . In terms of a probability distribution, this would correspond to a distribution  $d_\lambda(a, b|x, y)$  such that  $d_\lambda(a, b|x, y) = 1$  if and only if  $a = a_x$  and  $b = b_y$  as indicated by the vector  $\lambda$ , and  $d_\lambda(a, b|x, y) = 0$  otherwise. A behaviour, that is distributions  $p(a, b|x, y)$ , is local if and only if

$$p(a, b|x, y) = \sum_{\lambda} q_{\lambda} d_{\lambda}(a, b|x, y) \quad (32)$$

where the sum is taken over all  $|X|^{|A|}|Y|^{|B|}$  possible  $\lambda$ <sup>37</sup> where  $|A|$  and  $|B|$  denote the number of possible outputs for Alice and Bob, and

$$\forall \lambda, q_{\lambda} > 0, \text{ and } \sum_{\lambda} q_{\lambda} = 1. \quad (33)$$

We note that one can test whether such  $q_{\lambda}$  exists—i.e., whether the behaviour is local – using a linear program.<sup>43,44</sup> The dual of this linear program can be used to find a Bell inequality that certifies that a behaviour  $p(a, b|x, y)$  is *not* local.<sup>37</sup> One can easily adapt this linear program to search for an inequality that achieves a high violation. Specifically,

$$\begin{aligned} \text{maximise Violation} &= \sum_{x,y,a,b} s_{ab}^{xy} p(a, b|x, y) - S \\ \sum_{x,y,a,b} s_{ab}^{xy} d_{\lambda}(a, b|x, y) &\leq S, \forall \lambda \end{aligned} \quad (34)$$

$$0 \leq s_{ab}^{xy} \leq 1, \forall x, y, a, b, \quad (35)$$

where  $p(a, b|x, y)$  and  $d_{\lambda}(a, b|x, y)$  are given, and we optimise over  $s_{ab}^{xy}$  (see ref. 37 for details). Note that the second constraint means that for every LHVM we have a Bell inequality in which  $\beta_{\max} = S$ , and the difference  $V$  is precisely the violation we achieve when normalising the score functions to lie in the interval  $[0, 1]$ , which can be done without loss of generality.

It is clear from the discussion above that it can be to our advantage to search for a win/lose game, rather than a general game, as the  $P$  values for such games are sharper. This can be done by optimising over score functions in which  $s_{ab}^{xy} \in \{0, 1\}$ . This, however, is now an integer program<sup>45</sup> rather than a linear program,<sup>46</sup> which is in general NP-hard to solve.<sup>45,47</sup> Nevertheless, this may be feasible for the small number of inputs and outputs used in any experimental implementation, and heuristic methods exist.

Combining independent experiments

Suppose that a series of  $n$  experiments is run independently. Each experiment could correspond to completely different settings, Bell inequalities and so on. Associated with each experiment we obtain a series of  $P$  values corresponding to the probability that each of them was governed by an LHVM:  $(p_i)_{i=1}^n$ . In this situation, it is possible to take all the  $P$  values associated with each one of the individual experiments and obtain a

combined  $P$  value. One such method is Fisher's method.<sup>48,49</sup> With Fisher's method the combined  $P$  value is given by the tail probability of  $\chi_{2n}^2$ , a chi-squared distribution with  $2n$  degrees of freedom:

$$P\text{-value} = \Pr\left(\chi_{2n}^2 \geq -2 \log \prod_{i=1}^n p_i\right) \quad (36)$$

The right-hand side of this equation can be easily evaluated numerically. However, it can be shown that the tail probability of  $\chi_{2n}^2$  accepts the following closed expression:

$$\Pr(\chi_{2n}^2 \geq 2x) = e^{-x} \sum_{i=0}^{n-1} \frac{x^i}{i!} \quad (37)$$

where we can choose  $x = -\log \prod_{i=1}^n p_i$ .

However, we make no claim of optimality regarding the combined  $P$  value. There is rich literature on methods for combining  $P$  values,<sup>50</sup> and depending on the concrete situation a different choice should be made.

## Conclusions

We have shown how to derive (nearly) optimal  $P$  values for all Bell inequalities that can easily be applied to evaluate the data collected in experiments. A suitable Bell inequality can be found as outlined above; however, it would be interesting to combine this method with the numerical search for inequalities in.<sup>4,12,14</sup> The latter can adaptively find the best way to discriminate between LHVMs and theories like quantum mechanics that go beyond local-hidden variables that is asymptotically optimal but requires a significant amount of data to train.

We note that there exist many ways to extend the methods presented here to deal with specific situations at hand—for example, by conducting multiple experiments in succession and using data from prior runs to find more suitable Bell inequalities in the next instance.

We emphasise that the methods outlined here can be used to test other models than LHVMs. It is clear from the proof that only the winning probability in (13), or the expectation value (29), depends on the model to be tested. The argument that extends these bounds for a single trial to a bound on the  $P$  value for the entire experiment allowing arbitrary memory in the devices, however, does not depend on the model tested. In particular, this means that any theories that predict bounds of the form (13) and (29) are excluded with the same bound on  $P$  value. This also makes it apparent how one can extend the analysis to refute models that are more powerful than an LHVM. For example, Hall<sup>51</sup> defined and quantified interesting relaxations of an LHVM, with reduced free will, or where some amount of signalling is allowed. It is straightforward to adapt the analysis of<sup>51</sup> to derive bounds on (13) and (29) to subsequently obtain a  $P$  value for testing such extended models. Note that as Alice and Bob obtain an advantage by allowing models such as<sup>51</sup>—i.e., they are allowed more powerful strategies—they can achieve a higher score in the game. This implies that concrete scores will result in higher  $P$  values and lower confidence.

Furthermore, although we focussed the discussion on tests of Bell inequalities, our methods can also be applied to the study of certified randomness as in, refs 11,13,52 or more generally to tests of, e.g., non-contextual models that can be phrased as one player games.

## ACKNOWLEDGEMENTS

We thank Hannes Bernien, Peter Bierhorst, Andrew Doherty, Anaïs Dréau, Richard Gill, Peter Grünwald, Ronald Hanson, Bas Hensen, Jed Kaniewski, Laura Mančinska, Corsin Pfister, Tim Tamirniau, Thomas Vidick and Yanbao Zhang for discussions and/or comments on an earlier version of this manuscript. We also thank the referees for

their careful reading and suggestions. D.E. and S.W. are supported by STW, Netherlands, and an NWO VIDI Grant.

## COMPETING INTERESTS

The authors declare no conflict of interest.

## REFERENCES

- Bell, J. S. *Speakable and Unsayable in Quantum Mechanics: Collected Papers on Quantum Philosophy*. Cambridge Univ. Press, (2004).
- Gill, R. D. in *Proceedings of Foundations of Probability and Physics* Vol. 5, 179–206 (Vaxjö Univ. Press, 2003).
- Barrett, J., Collins, D., Hardy, L., Kent, A. & Popescu, S. Quantum nonlocality, Bell inequalities, and the memory loophole. *Phys. Rev. A* **66**, 042111 (2002).
- Zhang, Y., Glancy, S. & Knill, E. Asymptotically optimal data analysis for rejecting local realism. *Phys. Rev. A* **84**, 062118 (2011).
- Bierhorst, P. A rigorous analysis of the Clauser-Horne-Shimony-Holt inequality experiment when trials need not be independent. *Found. Phys.* **44**, 736–761 (2014).
- Gill, R. D. Accardi contra Bell (cum mundi): The impossible coupling. *IMS Lect. Notes Monogr. Ser.* **42**, 133–154 (2003).
- Van Dam, W., Gill, R. D. & Grunwald, P. D. The statistical strength of nonlocality proofs. *IEEE Trans. Inf. Theory* **51**, 2812–2835 (2005).
- Peres, A. Bayesian analysis of Bell inequalities. *Fortschr. Phys.* **48**, 531–535 (2000).
- Larsson, J. Å. & Gill, R. D. Bell's inequality and the coincidence-time loophole. *Europhys. Lett.* **67**, 707 (2004).
- Acín, A., Gill, R. D. & Gisin, N. Optimal Bell tests do not require maximally entangled states. *Phys. Rev. Lett.* **95**, 210402 (2005).
- Pironio, S. et al. Random numbers certified by Bell's theorem. *Nature* **464**, 1021–1024 (2010).
- Zhang, Y., Knill, E. & Glancy, S. Statistical strength of experiments to reject local realism with photon pairs and inefficient detectors. *Phys. Rev. A* **81**, 032117 (2010).
- Pironio, S. & Massar, S. Security of practical private randomness generation. *Phys. Rev. A* **87**, 012336 (2013).
- Zhang, Y., Glancy, S. & Knill, E. Efficient quantification of experimental evidence against local realism. *Phys. Rev. A* **88**, 052119 (2013).
- Gill, R. D. Statistics, causality and Bell's theorem. *Stat. Sci.* **29**, 512–528 (2014).
- Bierhorst, P. A robust mathematical model for a loophole-free Clauser-Horne experiment. *J. Phys. A* **48**, 195302 (2015).
- Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.* **49**, 1804 (1982).
- Wehner, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039 (1998).
- Tittel, W., Brendel, J., Gisin, N. & Zbinden, H. Long-distance Bell-type tests using energy-time entangled photons. *Phys. Rev. A* **59**, 4150 (1999).
- Rowe, M. A. et al. Experimental violation of a Bell's inequality with efficient detection. *Nature* **409**, 791–794 (2001).
- Matsukevich, D. N., Maunz, P., Moehring, D. L., Olmschenk, S. & Monroe, C. Bell inequality violation with two remote atomic qubits. *Phys. Rev. Lett.* **100**, 150404 (2008).
- Ansmann, M. et al. Violation of Bell's inequality in Josephson phase qubits. *Nature* **461**, 504–506 (2009).
- Scheidt, T. et al. Violation of local realism with freedom of choice. *Proc. Natl. Acad. Sci. USA* **107**, 19708–19713 (2010).
- Giustina, M. et al. Bell violation using entangled photons without the fair-sampling assumption. *Nature* **497**, 227–230 (2013).
- Christensen, B. G. et al. Detection-loophole-free test of quantum nonlocality, and applications. *Phys. Rev. Lett.* **111**, 130406 (2013).
- Pope, J. E. & Kay, A. Limited measurement dependence in multiple runs of a Bell test. *Phys. Rev. A* **88**, 032110 (2013).
- Bancal, J. D., Sheridan, L. & Scarani, V. More randomness from the same data. *New J. Phys.* **16**, 033011 (2014).
- Kofler, J. & Giustina, M. Requirements for a loophole-free Bell test using imperfect setting generators. Preprint at arXiv:1411.4787 (2014).
- Larsson, J. Å., Giustina, M., Kofler, J., Wittmann, B., Ursin, R. & Ramelow, S. Bell-inequality violation with entangled photons, free of the coincidence-time loophole. *Phys. Rev. A* **90**, 032107 (2014).
- Larsson, J. Å. Loopholes in Bell inequality tests of local realism. *J. Phys. A* **47**, 424003 (2014).
- Christensen, B. G. et al. Analysis of coincidence-time loopholes in experimental Bell tests. Preprint at arXiv:1503.07573 (2015).

32. Knill, E., Glancy, S., Nam, S. W., Coakley, K. & Zhang, Y. Bell inequalities for continuously emitting sources. *Phys. Rev. A* **91**, 032105 (2015).
33. Bentkus, V. On Hoeffding's inequalities. *Ann. Probab.* **32**, 1650–1673 (2004).
34. Clauser, J. F., M., Horne, A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880 (1969).
35. Clauser, J. F. & Horne, M. A. Experimental consequences of objective local theories. *Phys. Rev. D* **10**, 526 (1974).
36. Hensen, B. *et al.* Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686(2015).
37. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell non-locality. *Rev. Mod. Phys.* **86**, 419 (2014).
38. Cleve R., Hoyer P., Toner B., Watrous J. *19th IEEE Annual Conference on Computational Complexity* 236–249. (IEEE, Amherst, MA, USA, 2004).
39. Mermin, N. D. Extreme quantum entanglement in a superposition of macroscopically distinct states. *Phys. Rev. Lett.* **65**, 1838 (1990).
40. McDiarmid, C. On the method of bounded differences. *Surv. Combinator.* **141**, 148–188 (1989).
41. Collins, D., Gisin, N., Linden, N., Massar, S. & Popescu, S. Bell inequalities for arbitrarily high-dimensional systems. *Physical Rev. Lett.* **88**, 040404 (2002).
42. Shafer, G., Shen, A., Vereshchagin, N. & Vovk, V. Test martingales, Bayes factors, and p-values. *Stat. Sci.* **26**, 84–101 (2011).
43. Żukowski, M., Kaszlikowski, D., Batura, A. & Larsson, J.-A. Strengthening the Bell theorem: conditions to falsify local realism in an experiment. Preprint at quant-ph/9910058 (1999).
44. Kaszlikowski, D., Gnaniński, P., Żukowski, M., Miklaszewski, W. & Zeilinger, A. Violations of local realism by two entangled n-dimensional systems are stronger than for two qubits. *Phys. Rev. Lett.* **85**, 4418 (2000).
45. Faigle, U., Kern, W. & Still, G. *Algorithmic principles of mathematical programming* Vol. 24 (Springer Science & Business Media, 2013).
46. Boyd, S. & Vandenberghe, L. *Convex Optimization* (Cambridge Univ. Press, 2004).
47. Arora, S. & Boaz B. *Computational complexity: a modern approach*. (Cambridge University Press, 2009).
48. Fisher, R. A. *Statistical Methods for Research Workers* (Genesis Publishing Pvt Ltd, 1925).
49. Elston, R. C. On Fisher's method of combining p-values. *Biometr. J.* **33**, 339–345 (1991).
50. Loughin, T. M. A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.* **47**, 467–485 (2004).
51. Hall, M. J. W. Relaxed Bell inequalities and Kochen-Specker theorems. *Phys. Rev. A* **84**, 022102 (2011).
52. Colbeck, R. *Quantum And Relativistic Protocols For Secure Multi-Party Computation*, PhD thesis Univ. Cambridge, (2007).
53. Abellán, C. *et al.* Ultra-fast quantum randomness generation by accelerated phase diffusion in a pulsed laser diode. *Opt. Express* **22**, 1645–1654 (2014).
54. Abellán, C., Amaya, W., Mitrani, D., Pruneri, V. & Mitchell, M. W. Generation of fresh and pure random numbers for loophole-free Bell tests. *Physical review letters* **115**, 250403 (2015).
55. Pan, J.-W., Bouwmeester, D., Daniell, M., Weinfurter, H. & Zeilinger, A. Experimental test of quantum nonlocality in three-photon Greenberger-Horne-Zeilinger entanglement. *Nature* **403**, 515–519 (2000).
56. Zhao, Z., Yang, T., Chen, Y.-A., Zhang, A.-N., Żukowski, M. & Pan, J.-W. Experimental violation of local realism by four-photon Greenberger-Horne-Zeilinger entanglement. *Phys. Rev. Lett.* **91**, 180401 (2003).
57. Erven, C. *et al.* Experimental three-photon quantum nonlocality under strict locality conditions. *Nat. Photon.* **8**, 292–296 (2014).
58. Brassard, G., Broadbent, A. & Tapp, A. Recasting Mermin's multi-player game into the framework of pseudo-telepathy. Preprint at quant-ph/0408052 (2004).
59. Bell, J. S. Bertlmann's socks and the nature of reality. *J Phys Colloq.* **42**, C2–41 (1981).
60. Dada, A. C., Leach, J., Buller, G. S., Padgett, M. J. & Andersson, E. Experimental high-dimensional two-photon entanglement and violations of generalized Bell inequalities. *Nat. Phys.* **7**, 677–680 (2011).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016

Supplementary Information accompanies the paper on the *npj Quantum Information* website (<http://www.nature.com/npjqi>)