Melbourne Business School

From the Selected Works of Chris J. Lloyd

October, 2008

More powerful exact tests of equivalence

Chris Lloyd, Melbourne Business School

Available at: https://works.bepress.com/chris_lloyd/11/
ABSTRACT. In randomized clinical trials, it is often required to demonstrate that a new medical treatment is neither substantially worse nor better than a standard reference treatment. Formal testing of such ‘equivalence hypotheses’ is typically done by combining two one-sided tests (TOST). A quite different strand of research has demonstrated that maximizing P-values over nuisance parameters produces optimal tests (Rohmel and Mansmann (1999) and Lloyd (2008a)). In this paper we point out that, even if the one-sided tests are exact and optimal, the TOST will generally be conservative and requires a further adjustment to remove this conservatism. The appropriate procedure is suggested and illustrated with an application to the difference of response rates from a binary matched pairs design. Based on an extensive numerical study, we recommend a particular version of the TOST based on the estimation followed by maximization as a more powerful alternative to the known standard TOST procedures.

key words and phrases TOST, nuisance parameter

1 Introduction

In randomized clinical trials (RCT’s), the term equivalence often implicitly refers to demonstrating that a new experimental treatment (the test) is not substantially worse than a standard known alternative (the active control), see ICH E9. A more precise term for that would be noninferiority, which commonly involves a one-sided statistical procedure, either one of tests or confidence limits. However, sometimes it is explicitly required to demonstrate that a new treatment is neither substantially worse nor better than a standard reference treatment. For instance, demonstrating that a new generic drug is equivalent, neither worse nor better, than a known branded drug of the same formulation can greatly accelerate the drug approval process. On the other hand, the fact that the generic substitute is more effective than its branded counterpart can raise a series of questions, e.g. about potentially more severe adverse effects, and thus force an exceptionally long and costly full development. One more example is
related to comparison of antibiotics. In particular, excessive exposure to antibacterial agents can increase the disruption of the normal flora of human body, and therefore a significantly greater effect of the new antibacterial treatment as compared to a known active control is undesirable. Finally, the area where equivalence tests are in common use is so-called bioequivalence trials, trials intended to show that the new treatment is absorbed into the blood at approximately the same rate as the active control, with significant deviations in either direction being undesirable.

Let $\theta$ be the difference between the response probability for the two treatments and $\delta$ be a so-called noninferiority margin, normally selected as the smallest clinically important difference between the test and active control treatments. We are interested in simultaneously testing hypotheses of the form

$$H_L^0 : \theta \leq -\delta \ vs. \ H_L^1 : \theta > -\delta,$$

(1)
a test involving the lower margin, hereafter referred to as a lower test $L$, and

$$H_U^0 : \theta \geq \delta \ vs. \ H_U^1 : \theta < \delta,$$

(2)
a test involving the upper margin, hereafter referred to as an upper test $U$. We are interested in the combination of $L$ and $U$ given by

$$H_E^0 : H_L^0 \cup H_U^0 \ vs. \ H_E^1 : H_L^1 \cap H_U^1,$$

(3)
which corresponds to a test of equivalence. This test is an example of an intersection-union test (IUT), which leads to the two one-sided test (TOST) approach, see Berger and Hsu (1996). The alternative equivalence hypothesis $H_E^1$ says that $\theta$ is within the open interval $(-\delta, \delta)$. In practice, power of an equivalence test tends to be rather low and it is important to employ a testing procedure that maximizes it as much as possible.

The objective of our study is to present a more powerful exact unconditional statistical testing strategy for demonstration of equivalence of two treatments. The suggested strategy leads to tests which are less conservative and more powerful than the traditional TOST alternatives commonly employed in practice. The suggested testing strategy is applied for testing the equivalence of response rates from a binary matched pairs design.
2 The binary matched pairs

In order to account for population heterogeneity, it is common to conduct matched-pairs design clinical trials where subjects are either matched according to a specific covariate with further randomization to test and control groups, e.g. in intervention studies, or each assigned to receive the sequence of test and control in one of two possible orders, e.g. in cross-over design trials, see Hsueh et al. (2001) and Agresti and Min (2005) for some standard methods of analysis.

In general, we have \( n \) individuals who have two binary responses measured. Denote by \( X_{jk} \) the number of responses \( jk \in \{11, 01, 10, 11\} \) with corresponding probabilities \( \pi_{jk} \). We are interested in the parameter \( \theta = \pi_{01} - \pi_{10} \), which is the probability of response from method 2 (the test) minus the probability of response for method 1 (the control).

Provided individuals within each arm respond independently, the joint distribution of \( \{X_{jk}\} \) is multinomial but can be expressed as a product of three binomial factors, namely

\[
B(t; n, \phi)B(x_{01}; t, \eta)B(x_{11}; n - t, \psi),
\]

where \( B(x; n, p) \) is the probability of a binomial with parameters \( (n, p) \) equaling \( x \) and \( t = x_{01} + x_{10} \) is the number of so-called “discordant” pairs. The parameter \( \phi = \pi_{01} + \pi_{10} \) is the probability of a discordant pair, while \( \eta = (\theta + \phi)/(2\phi) \) is the probability of a response (01) in favor of the treatment, conditional on the response being discordant. The parameter \( \psi = \pi_{11}/(\pi_{00} + \pi_{11}) \) and it is commonly agreed on the basis of sufficiency and/or conditionality principles that the factor involving this parameter has no relevance to inference on \( (\theta, \phi) \). Denoting \( x_{01} \) by \( x \), the data comprise \( y = (x, t) \) and the likelihood is

\[
L(\theta, \phi; y, n) \propto \phi^t(1 - \phi)^{n-t}\eta^x(1 - \eta)^{t-x} \propto (1 - \phi)^{n-t}(\theta + \phi)^x(\phi - \theta)^{t-x}
\]

defined over the parameter space \( \Omega := \{ (\theta, \phi) : 0 \leq |\theta| \leq \phi \leq 1 \} \). The maximum likelihood (ML) estimate of \( \theta \) is \( \hat{\theta} = (2x - t)/n \) and has asymptotic variance \( \sigma^2/n \), where \( \sigma^2 = (\phi - \theta^2)/n \). The profile ML estimate \( \hat{\phi}_\theta \) of \( \phi \) when \( \theta \) is assumed to be known is given by the larger solution of the quadratic equation

\[
\phi^2 - \phi(\hat{\phi} - \hat{\theta}\theta) + \hat{\theta}\theta - (1 - \hat{\phi})\theta^2 = 0,
\]

which is necessarily within the interval \([|\theta|, 1]\). The unrestricted ML estimate of \( \phi \) is \( \hat{\phi} = t/n \), which is equivalent to the profile estimate when \( \theta = 0 \).
Standard likelihood theory leads to the Wald-type test statistics \( \sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma} \) with approximate standard normal distribution. Using the restricted ML estimator \( \hat{\phi}_0 - \theta^2 \) of \( \sigma^2 \) under the null hypothesis, leads to the statistic of Nam (1997) and Tango (1998) who independently showed that this is identical to Rao’s score statistic

\[
R(y; \theta) = \sqrt{n}(\hat{\theta} - \theta)/\sqrt{\hat{\phi}_0 - \theta^2}.
\]

Other estimators of \( \sigma \) in the Wald statistic have been suggested by Lu and Bean (1995) and Hsueh et al. (2001) but are not considered in our study for reasons described in Hsueh et al. (2001) and Liu et al. (2002). All Wald-type statistics reduce to the statistics of McNemar (1947) when \( \theta = 0 \).

An alternative to \( R(y; \theta) \) first suggested by Lloyd and Moldovan (2008) is the signed root likelihood ratio (LR) statistic that has the form

\[
L(y; \theta) = \text{sign}(\hat{\theta} - \theta)\text{LR}(y, \theta)^{1/2},
\]

where \( \text{LR}(y, \theta) = 2\log \left( L(\hat{\theta}, \hat{\phi})/L(\theta, \hat{\phi}_0) \right) \) with the conventions that both \( 0 \log 0 \) and \( \text{sign}(0) \) equal zero. Both the score and LR statistics are asymptotically normal when \( n \) is large and when \( \phi \) is not near the boundary of \( \theta, 1 \).

Let \( S(y; \theta) \) denote a generic test statistic for testing \( \theta \), either the score or LR statistics in our case. For testing against the upper margin \( \delta \), large negative values of \( S(y; \delta) \) lead to rejection of the null hypothesis and an approximate \( P \)-value based on the approximating normal distribution is \( P_U(y; \delta) = \Phi(S(y; \delta)) \). There is an inherent symmetry in the matched pairs model that implies a symmetry between tests \( U \) and \( L \). If a particular test statistic \( S(y, \delta) \) is optimal for \( U \), it is also optimal for \( L \) and the corresponding \( P \)-value is \( P_L(y; \delta) = P_U(y'; \delta) \) with \( y' = (t - x, t) \).

Two-sided equivalence tests are typically based on combining the two one-sided tests (TOST) of \( L \) and \( U \) with the same target size, see Berger and Hsu (1996, Theorem 1). Formally, the TOST rejects the null hypothesis when the maximum of the two one sided \( P \)-values is less than \( \alpha \). One can think of this as a transformation \( T \) of the original \( P_U \)-value

\[
T : P_U(y; \delta) \rightarrow P_E(y; \delta) := \max\{P_U(y; \delta), P_U(y'; \delta)\}.
\]

Under some circumstances, see Berger and Hsu (1996, Theorem 2), the size of the TOST will achieve the upper bound \( \alpha \) so that combining efficient one-sided tests automatically leads to an efficient equivalence test. However, the conditions for this do not hold for
matched pairs as well as for most discrete data models. Even beginning with an efficient one-sided test, the corresponding TOST will generally be conservative. Further in the study, we demonstrate how to remove this conservatism.

3 Exact tests and TOST

As pointed out above, it is common to use a TOST procedure for testing equivalence. This procedure involves two one-sided tests, which can be expressed in terms of rejecting the null if a P-value $P(y; \delta)$ is smaller than the chosen significance level $\alpha$. The exact P-value

$$\pi(y; \delta, \phi) := \Pr(P(Y; \delta) \leq P(y; \delta); \delta, \phi)$$

depends on $\phi \in (|\delta|, 1)$. This probability is calculated using the binomial distribution of $T$ with parameters $(n, \phi)$ and the conditional binomial distribution of $X$ given $T$ with parameters $(t, \eta)$, as indicated in Section 2. There are three main methods of controlling for the effect of $\phi$. Each method is based on consideration of $\pi(y; \delta, \phi)$ as a function of $\phi$, which is called the significance profile, and we express them below in terms of transformation of an approximate P-value to exact form. This theory is identical for both one- and two-sided tests and the three P-values defined below are guaranteed to satisfy the characteristic property of a valid P-value, namely

$$\sup_{\phi} \Pr(P(Y; \delta) \leq \alpha; \delta, \phi) \leq \alpha. \quad (7)$$

Full Maximization: $M$ P-values. Bickel and Doksum (1977, p.168) define the P-value to be the supremum of the significance profile. One can think of this as a transformation

$$M : P(y; \delta) \rightarrow P^*(y; \delta) := \sup_{\phi} \{\pi(y; \delta, \phi), \phi \in (|\delta|, 1)\}. \quad (8)$$

It can be shown that $P^*(Y; \delta)$ is as small as possible amongst valid P-values that are non-decreasing functions of the original statistic $P(Y; \delta)$, see Rohmel and mansmann (1999) and Lloyd (2008a). It also satisfies (7) with the inequality replaced by a corresponding equality. This is an extremely powerful result and means that, in principle at least, all tests base based on maximised P-values.

Partial Maximization: $B$ P-values. Berger and Boos (1994) suggested the P-value

$$B : P(y; \delta) \rightarrow P_\gamma(y; \delta) := \sup_{\phi} \{\pi(y; \delta, \phi) : \phi \in C_\gamma\} + \gamma, \quad (9)$$
where $C_\gamma$ is a $100(1 - \gamma)$% confidence interval for $\phi$ under the null. This methodology has been applied in several recent papers, see for instance Berger and Sidik (2003) and Sidik (2003). Proponents of this approach acknowledge that dependence of results on the choice of $\gamma$ can be extreme, notwithstanding the general recommendation by Berger and Sidik that $\gamma$ be small. It is essential, therefore, that the confidence interval error $\gamma$ be strictly controlled. Sidik (2003) suggested using a Clopper-Pearson interval for $\phi$ intersected with the null interval $(|\delta|, 1)$. Clearly, $B$ P-values depend on the choice of $\gamma$, while no such subjective choices are required for $M$ P-values. Partially maximised P-values are necessarily not minimal, but it is easy to show that the degree of conservatism is less than $\gamma$ for all $y$.

*Estimation Followed by Maximization: $E+M$ P-values.* An alternative and much older approach is to replace the nuisance parameter by its estimate under the null. This $E$-step produces the statistic

$$
E : P(y; \delta) \rightarrow \hat{P}(y; \delta) := \pi(y; \delta, \hat{\phi}_\delta).
$$

While $\hat{P}(y; \delta)$ is not a valid P-value, it is valid after a further $M$-step given by (8). This involves computing the significance profile of the statistic $\hat{P}(y; \delta)$ which is different from that of $P(Y; \delta)$ and is typically much better behaved. Lloyd and Moldovan (2008) demonstrated that applied to difference between binary matched pairs one-sided $E+M$ tests are generally more powerful than $M$ and $B$ exact alternatives.

In terms of transformations introduced above, the traditional combinations of two one-sided tests in the TOST procedure are $MT$, $BT$ or $EMT$, which lead to conservative tests. We suggest removing this conservatism by an additional exact transformation, either $M$ or $B$, after the $T$-step. Luckily, we can show that $M$ and $B$ need only be applied only once at the end, because the results depend only on the ordering of the generating statistic, see Appendix. Thus, the new exact equivalence P-values of the form $TM$, $TB$ and $ETM$ require the same computation effort as current methods. In the next Section, we report the results of the numerical study comparing the range of alternative TOST procedures on the basis of their size and rejection power.

## 4 Numerical study

In this section, we compare the traditional TOST procedures given by $MT$, $BT$ and $EMT$ with the new TOST strategies based on $TM$, $TB$ and $ETM$ transforms. All our comparisons are based on a full enumeration of the competing P-values for all
(n + 1)(n + 2)/2 possible data sets. The comparisons we used are based on the actual size and the power of each test with \( \delta = 0.10 \), both computed using

\[
\beta(\theta, \phi) = \Pr(P(Y) \leq \alpha; \theta, \phi),
\]

where the P-value \( P(Y) \) does depend on both \( \alpha \) and \( \delta \). The achieved size of the test is measured by \( \sup_{\phi} \beta(\delta, \phi) \).

Lloyd (2008b) demonstrates that \( \phi \) must be less than \( \theta^* = 0.5 \sqrt{1 + \theta^2} \) assuming only that matching induces a non-negative correlation between the binary responses. When \( \delta = 0 \) this limit is \( \theta^* = 0.5 \) and when \( \delta = 0.10 \) it equals \( \theta^* = 0.5025 \). We therefore calculate the achieved size over the range \( \phi \in (0, 0.5025) \). For the test power, we will look at the most interesting point of the alternative parameter value, namely \( \theta = 0 \). Specifically, we look at the power profile \( \beta(0, \phi) \) where \( \phi \in (0, 0.5) \), keeping in mind that in many typical matched studies the value of \( \phi \) will likely be smaller than 0.5 and the achieved power larger than the reported summary values. The power profile typically has high values, often close of equal to 1, when \( \phi \) is close to 0 and a much lower values when \( \phi \) is close to \( \phi = 0.5 \). We will give a typical power profile plot below, but will end up using the mean value of corresponding power functions.

Tables 1 and 2 below report the results of achieved size and power. All the results are based only on \( \delta = 0.10 \). We found that tests involving another typical noninferiority margin \( \delta = 0.05 \) are impractical due to discreteness of the sample space, which becomes intolerable for such a narrow equivalence range as \((-0.05, 0.05)\).

As can be seen from Table 1, the achieved size never exceeds nominal which is due to application of the exact transforms \( \mathcal{M} \) and \( \mathcal{B} \). The six tests are grouped in pairs according the order of the maximization step application. In particular, \( \mathcal{M} \) and \( \mathcal{B} \) steps are applied either before the \( T \)-step, as it is currently in practice, or after the \( T \)-step, according to our suggestion. As could be expected, in many cases the maximization step applied after the \( T \)-step reduces the conservatism of the traditional exact tests \( MT, BT \) and \( EMT \) bringing the achieved size closer to nominal. In most cases, the improvement in the achieved size also leads to an increased test power, as evident from Table 2. Note that the achieved size and power results of the suggested new tests \( TM, TB \) and \( ETM \) never worse and in many cases better than the corresponding results of their traditional counterparts \( MT, BT \) and \( EMT \). Among the new tests, the test based on estimation followed by maximization \( ETM \) from the score statistic has the highest average power. This result is consistent with Lloyd and Moldovan (2008), who also suggested the test based on estimation followed by maximization from the score
Table 1: Comparison of six alternative equivalence tests of nominal sizes 5% and 10%. Entries are achieved test size \( \sup_\phi \beta(\delta, \phi) \) over the interval \( \phi \in [0, \theta^*] \), \( \theta^* = 0.5\sqrt{1 + \theta^2} \) with \( \theta = \delta \). The noninferiority margin \( \delta = 0.10 \).

<table>
<thead>
<tr>
<th>Test</th>
<th>( \alpha = 0.05 )</th>
<th>( n = 50 )</th>
<th>( n = 75 )</th>
<th>( n = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>LR</td>
<td>Score</td>
<td>LR</td>
</tr>
<tr>
<td>MT</td>
<td>0.0442 0.0442</td>
<td>0.0334 0.0315</td>
<td>0.0453 0.0384</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td>0.0445 0.0445</td>
<td>0.0334 0.0315</td>
<td>0.0453 0.0384</td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td>0.0442 0.0442</td>
<td>0.0334 0.0315</td>
<td>0.0453 0.0384</td>
<td></td>
</tr>
<tr>
<td>TB</td>
<td>0.0445 0.0445</td>
<td>0.0334 0.0315</td>
<td>0.0453 0.0384</td>
<td></td>
</tr>
<tr>
<td>EMT</td>
<td>0.0442 0.0442</td>
<td>0.0440 0.0440</td>
<td>0.0482 0.0482</td>
<td></td>
</tr>
<tr>
<td>ETM</td>
<td>0.0445 0.0445</td>
<td>0.0440 0.0440</td>
<td>0.0483 0.0483</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \alpha = 0.10 )</td>
<td>Score</td>
<td>LR</td>
<td>Score</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LR</td>
<td>Score</td>
<td>LR</td>
</tr>
<tr>
<td>MT</td>
<td>0.0611 0.0605</td>
<td>0.0903 0.0745</td>
<td>0.0903 0.0801</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td>0.0611 0.0605</td>
<td>0.0903 0.0745</td>
<td>0.0903 0.0801</td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td>0.0611 0.0605</td>
<td>0.0903 0.0745</td>
<td>0.0908 0.0885</td>
<td></td>
</tr>
<tr>
<td>TB</td>
<td>0.0611 0.0605</td>
<td>0.0903 0.0745</td>
<td>0.0908 0.0885</td>
<td></td>
</tr>
<tr>
<td>EMT</td>
<td>0.0768 0.0768</td>
<td>0.0915 0.0915</td>
<td>0.0983 0.0908</td>
<td></td>
</tr>
<tr>
<td>ETM</td>
<td>0.0908 0.0769</td>
<td>0.0915 0.0915</td>
<td>0.0983 0.0983</td>
<td></td>
</tr>
</tbody>
</table>

It can be noticed that sometimes alternative testing strategies have identical power for tests of nominal size 0.05 but different power for tests of nominal size 0.10. To clarify this point, the left panel of Figure 1 plots ETM P-values against TM P-values computed from the LR statistic for \( n = 50 \). It can be seen that the 5% tests generated from these P-values are identical in terms of rejecting the null. The corresponding 10% tests already have rejection disagreement in favor of the ETM test. The right panel of Figure 1 gives the two power profiles for the 10% tests. The dominance of the ETM test (the solid line) in terms of power is obvious and these power profiles are typical for the rest of tests.

5 Conclusion

Two-sided equivalence test often have low power due to the relatively narrow range of the rejection region of the parameter space given by \((-\delta, \delta)\). Hence, in practice it is important to use the testing procedure with maximum power. We have pointed out that maximisation is an essential step in designing efficient tests and, that for two-sided equivalence tests, this maximisation should be done after combining the two one-sided
Table 2: Comparison of six alternative equivalence tests of nominal sizes 5% and 10%. Entries are mean of power profile $\beta(\theta, \phi)$ over the interval $\phi \in [0, \theta^*]$, $\theta^* = 0.5 \sqrt{1 + \theta^2}$ at the alternative $\theta = 0$. The noninferiority margin $\delta = 0.10$.

<table>
<thead>
<tr>
<th>Test $\alpha = 0.05$</th>
<th>$n = 50$</th>
<th>$n = 75$</th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>LR</td>
<td>Score</td>
</tr>
<tr>
<td>MT</td>
<td>0.1551</td>
<td>0.1551</td>
<td>0.2288</td>
</tr>
<tr>
<td>TM</td>
<td>0.1658</td>
<td>0.1658</td>
<td>0.2288</td>
</tr>
<tr>
<td>BT</td>
<td>0.1551</td>
<td>0.1551</td>
<td>0.2336</td>
</tr>
<tr>
<td>TB</td>
<td>0.1658</td>
<td>0.1658</td>
<td>0.2383</td>
</tr>
<tr>
<td>EMT</td>
<td>0.1551</td>
<td>0.1551</td>
<td>0.2561</td>
</tr>
<tr>
<td>ETM</td>
<td>0.1658</td>
<td>0.1658</td>
<td>0.2561</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test $\alpha = 0.10$</th>
<th>$n = 50$</th>
<th>$n = 75$</th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>LR</td>
<td>Score</td>
</tr>
<tr>
<td>MT</td>
<td>0.1999</td>
<td>0.1903</td>
<td>0.3817</td>
</tr>
<tr>
<td>TM</td>
<td>0.1999</td>
<td>0.1903</td>
<td>0.3817</td>
</tr>
<tr>
<td>BT</td>
<td>0.1999</td>
<td>0.1903</td>
<td>0.3934</td>
</tr>
<tr>
<td>TB</td>
<td>0.1999</td>
<td>0.1903</td>
<td>0.4099</td>
</tr>
<tr>
<td>EMT</td>
<td>0.2301</td>
<td>0.2301</td>
<td>0.4026</td>
</tr>
<tr>
<td>ETM</td>
<td>0.2576</td>
<td>0.2383</td>
<td>0.4136</td>
</tr>
</tbody>
</table>

In particular, we note that the traditional combination of exact one-sided tests through the TOST procedure leads to equivalence tests which are often unnecessarily conservative. The application of maximization over the nuisance parameter applied after the TOST transform removes this conservatism. We proposed a new equivalence test based on binary matched pairs which is more powerful among all known alternatives. The numerical study suggested that the test based on estimation followed by maximization (i.e. ETM) from the score statistic is the most powerful among the alternative tests considered and we recommend it to practitioners.

References.


Figure 1: **Power profiles.** Left. ETM P-values versus TM P-values for LR statistics with $n = 50$ and $\delta = 0.10$. Right. Power profiles $\beta(0, \phi)$ for $\alpha = 0.10$.


Tango, T. (1998) Equivalence test and confidence interval for the difference in propor-
tions for the paired-sample design. *Statistics in Medicine* 17, 891-908.

**Appendix.** *Ordering and the TOST transform.*

Suppose that we begin with two P-values $P_1$ and $P_2$ that are ordering compatible in the sense that

$$P_1(y_1) \geq P_1(y_2) \Rightarrow P_2(y_1) \geq P_2(y_2) \quad (11)$$

Let $P_{E_i}$ be the corresponding TOST P-values i.e.

$$P_{E_i}(y) := \max\{P_j(y), P_j(y')\}.$$

We aim to show that $P_{E_1}$ and $P_{E_2}$ are ordering compatible, i.e. that

$$P_{E_1}(y_1) \geq P_{E_1}(y_2) \Rightarrow P_{E_2}(y_1) \geq P_{E_2}(y_2)$$

which we will achieve by showing the contrapositive, namely

$$P_{E_2}(y_2) < P_{E_2}(y_1) \Rightarrow P_{E_1}(y_2) < P_{E_1}(y_1).$$

Supposing the left-hand side, and concentrating on whether $P_{E_2}(y_1)$ is achieved at $y_1$ or $y_1'$ there are two (possibly overlapping) cases.

**Case 1.** $P_{E_2}(y_1) = P_2(y_1)$. It follows that $P_{E_2}(y_2) < P_2(y_1)$ and so

$$P_2(y_2) < P_2(y_1) \& P_2(y_2') < P_2(y_1).$$

Using the contrapositive of (11) it follows that

$$P_1(y_2) < P_1(y_1) \& P_1(y_2') < P_1(y_1).$$

Hence

$$P_{E_1}(y_2) < P_1(y_1) \leq P_{E_1}(y_1).$$

**Case 2.** $P_{E_2}(y_1) = P_2(y_1')$. It follows that $P_{E_2}(y_2) < P_2(y_1')$ and so

$$P_2(y_2) < P_2(y_1') \& P_2(y_2') < P_2(y_1').$$

Using the contrapositive of (11) it follows that

$$P_1(y_2) < P_1(y_1') \& P_1(y_2') < P_1(y_1').$$

Hence

$$P_{E_1}(y_2) < P_1(y_1') \leq P_{E_1}(y_1).$$