# Extracting Knowledge Using Decision Principles: Proper Scoring Rules Reconsidered from the Perspective of Modern Decision Theories



Theo Offerman Joep Sonnemans Gijs van de Kuilen Peter P. Wakker

Review of Economic Studies 76, 1461–1489.

This is 1st lecture
in 1st ADT conference
(International Conference on Algorithmic Decision Theory),
Venice, Italy
Welcome all of us to a new community, joining many European forces!

I asked participants about central topic here:
Multicriteria optimization.
"Criteria" can be:

1. Beliefs or well-being of different people (aggregating over people);
2. Resolutions of uncertainty (decision making under uncertainty);
3. Payoff at different time points (dynamic/ sequential decision making).
4. And so on!

Most people here work algorithmically: find optimal solution.
I work in behavioral decision theory: what do people really do, empirically? Where deviate from optimum? So, where to improve actual decisions?

What better to start ADT with here in Italy than:
The nicest multi-criteria optimization problem ever invented?
You will see …
Was invented by Bruno de Finetti!

Topic this lecture: Uncertainty in optimization.
How extract knowledge using decision making theory?
How measure subjective belief of others (such as of experts, say weather forecasters)?

Say about the uncertain event:
***D*** = Next president of US will be ***D***emocrat;
or: Will client repay loan?
Major part of lecture: measure belief of, say, you in ***D***.

We first consider another application:
grading students.
`We measure their belief in ***D***.

Say, you grade a multiple choice exam in geography to test students' knowledge about
Statement ***D***: Capital of North Holland = Amsterdam.

| | Reward: if ***D*** true | if not ***D*** true |
|---|---|---|
| ***D*** | 1 | 0 |
| not-***D*** | 0 | 1 |

Problem: Correct answer does not completely identify student's knowledge.
Some correct answers, and high grades, are due to luck. There is noise in the data.

## Attempted solution:

Find r such that student indifferent between:

| | Reward: if ***D*** true | if not ***D*** true |
|---|---|---|
| ***D*** | 1 | 0 |
| partly know ***D***, to degree r | r | r |

Then r = P(***D***). (Assuming expected value maximization …)

How measure r?

1. Observe many binary choices between such options. Popular in decision theory. Problem: too crude and time consuming.
2. Just ask student what r is. Problem: why would they tell the truth??

I now promise a perfect way out:
de Finetti's dream-optimization problem;
a very clever two-criteria continuous optimization problem.
Will exactly identify state of knowledge of each student, no matter what it is.
Takes little time; no more than multiple choice.
Rewards students fairly, with little noise.
Best of all worlds. Here it is:

For all conceivable degrees of knowledge.

---

Student can choose reported probability r for ***D*** from the [0,1] continuum, as follows:

| | Criterium 1 | Criterium 2 |
|---|---|---|
| | Reward: if ***D*** true | if not ***D*** true |
| r=1: (***D*** = sure!?) | 1 | 0 |
| r: degree of belief in ***D*** (?) | $1 - (1-r)^2$ | $1-r^2$ |
| r=0.5: (have no clue!?) | 0.75 | 0.75 |
| r=0: (not-***D*** is sure!?) | 0 | 1 |

Claim: Under "subjective expected value," optimal reported probability r

=

true subjective probability p.

# Proof of claim.

To help memory:

| r: | Reward: if *D* true | if not *D* true |
|---|---|---|
| degree of belief in *D* | $1 - (1-r)^2$ | $1-r^2$ |

p true probability; r reported probability.

Optimize EV = $p(1 - (1-r)^2) + (1-p)(1-r^2)$.

1st order optimality:

$$2p(1-r) - 2r(1-p) = 0.$$

$$r = p!$$

Easy in algebraic sense.
Conceptually: !!! Wow !!!

Can read minds of people!
Incentive compatible ... Many implications ...
de Finetti (1962) and Brier (1950) were the first neuro-scientists.
They invented the nicest multi-criteria optimization problem ever!
Useful in many domains.

# "Bayesian truth serum" (Prelec, Science, 2005). Superior to elicitations through preferences [illegible]. Superior to elicitations through indifferences ~.

Widely used: Hanson (Nature, 2002), Prelec (Science 2005). In accounting (Wright 1988), Bayesian statistics (Savage 1971), business (Stael von Holstein 1972), education (Echternacht 1972), finance (Shiller, Kon-Ya, & Tsutsui 1996), medicine (Spiegelhalter 1986), psychology (Liberman & Tversky 1993; McClelland & Bolger 1994), experimental economics (Nyarko & Schotter 2002).

Remember: based on expected value; in 2009 …!?

We bring

- realism (of prospect theory) to proper scoring rules;
- the beauty of proper scoring rules to prospect theory and studies of ambiguity.

# Survey

Part I. Deriving reported prob. r from theories (different goal functions):

- expected value;
- expected utility;
- nonexpected utility for probabilities;
- imprecize/unknown probabilities.

Part II. Deriving theories from observed r.
In particular: Derive beliefs/ambiguity attitudes. Will be surprisingly easy.

Part III. Implementation in an experiment.

# Part I. Deriving r from Theories (EV, and then 3 deviating goal functions).

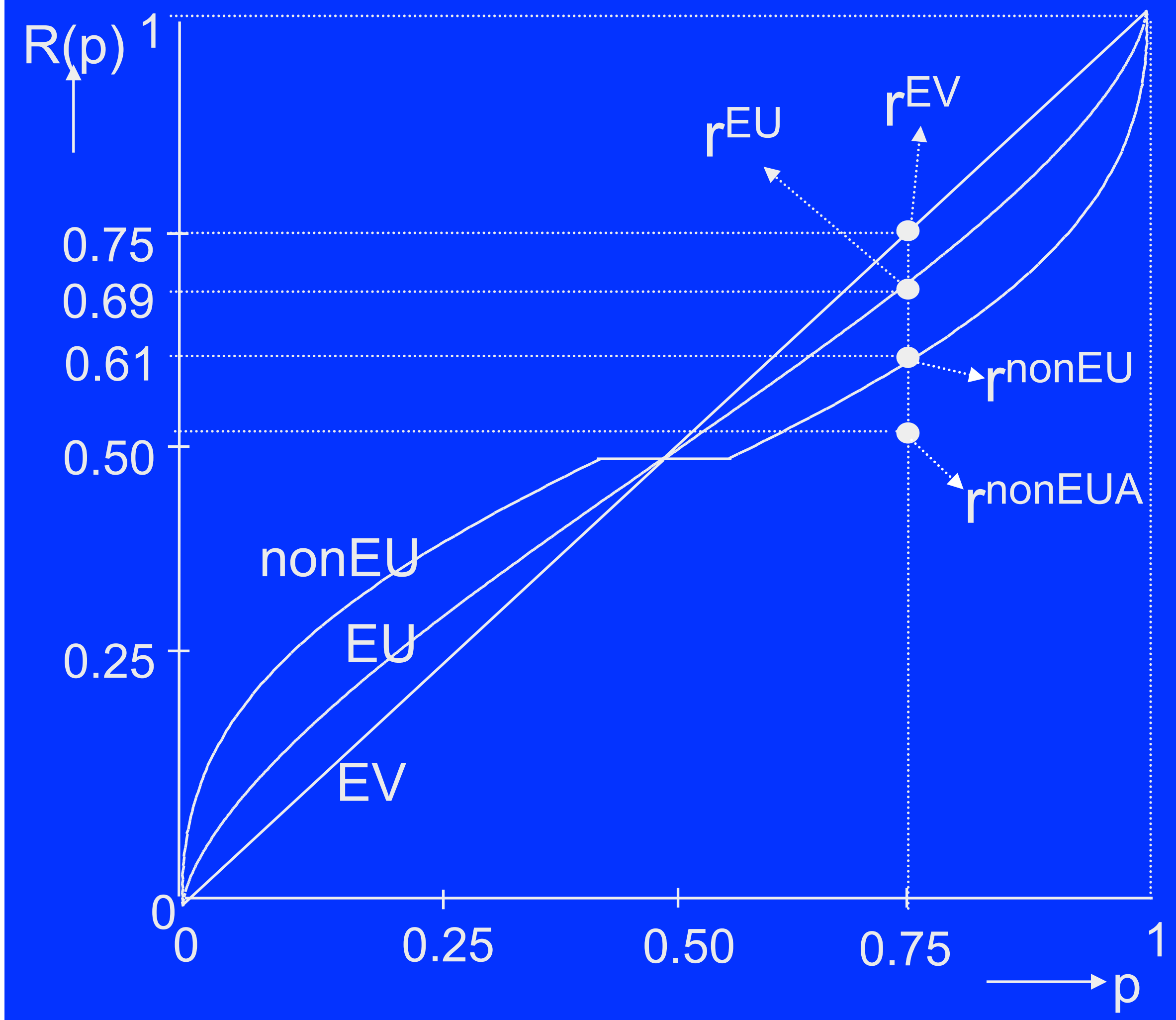
Event ***D***: Next president US is ***D***emocrat.
not-***D***: Next president is not democrat.

We quantitatively measure your subjective belief in this event
(subjective probability?; imprecize probability?),
i.e. how much you believe in ***D***.

Say your subjective probability of ***D*** = 0.75.

EV:

Then your optimal $r_D$ = 0.75.




Reported probability R(p) = $r_D$ as function of true probability p, under:

(a) expected value (EV);

(b) expected utility with U(x) = √x (EU);

(c) nonexpected utility for known probabilities, with U(x) = $x^{0.5}$ and with w(p) as common;

$r^{nonEUA}$: nonexpected utility for unknown probabilities ("Ambiguity").

next p.

go to p. 18, Example EU

go to p. 22, Example nonEU

go to p. 26, Example nonEUA

So far we assumed EV
(as in every application of proper scoring rules;
as in no modern risk-ambiguity theory ...)

Deviation 1 from EV: EU with U nonlinear

Now optimize
$pU(1 - (1- r)^2) + (1 - p)U(1 - r^2)$

r = p need no more be optimal.

**Theorem.** Under expected utility with true probability p,

$$r = \frac{p}{p + (1-p)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}$$

Reversed (and explicit) expression:

$$p = \frac{r}{r + (1-r)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}$$

How bet on ***D***? [Expected Utility].

EV: $r^{EV} = 0.75$.

Expected utility, $U(x) = \sqrt{x}$:

$r^{EU} = 0.69$.

You now bet less on ***D***. Closer to safety (50-50) (Winkler & Murphy 1970).

go to p. 15 with figure of R(p)

# Deviation 2 from EV: nonexpected utility for probabilities (Allais 1953, Machina 1982, Kahneman & Tversky 1979, Quiggin 1982, Gul 1991, Luce & Fishburn 1991, Tversky & Kahneman 1992; Birnbaum 2005; survey: Starmer 2000)

For two-gain prospects, virtually all those theories are as follows:

For $r \geq 0.5$, nonEU(r) =

$w(p)U(1 - (1-r)^2) + (1-w(p))U(1-r^2)$.

$r < 0.5$, symmetry, etc.

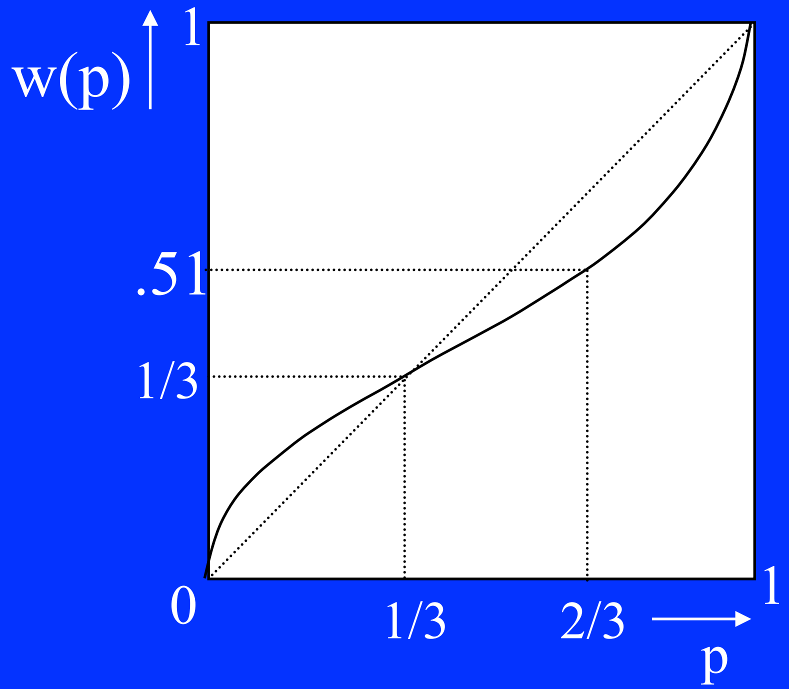



Figure. The common weighting function w.

$w(p) = \exp(-(-ln(p))^{\alpha})$ for $\alpha = 0.65$.

$w(1/3) \approx 1/3$;

$w(2/3) \approx .51$

**Theorem.** Under nonexpected utility with true probability p,

$$r = \frac{w(p)}{w(p) + (1-w(p))\,\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}$$

**Reversed (explicit) expression:**

$$p = w^{-1}\left(\frac{r}{r + (1-r)\,\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\right)$$

## How bet on ***D*** now? [nonEU with probabilities].

EV: $r^{EV} = 0.75$.

EU: $r^{EU} = 0.69$.

Nonexpected utility, $U(x) = \sqrt{x}$,

$$w(p) = \exp(-(-\mathit{ln}(p))^{0.65}).$$

$r^{nonEU} = 0.61$.

You bet even less on ***D***. Again closer to 50-50 safety.

go to p. 15, with figure of R(p)

# Deviation 3 from EV: Ambiguity

(unknown probabilities).

How deal with unknown/impricise probabilities?

Even have to give up probabilities ("Bayesian beliefs").

Instead of additive beliefs p = P(***D***), nonadditive beliefs B(***D***):

- Imprecise probabilities;
- upper/lower probabilities;
- Dempster&Shafer belief functions;
- Tversky& Koehler support functions;
- Zadeh-Morufushi/Sugeno fuzzy measures.

Virtually all decision models existing today:
For $r \geq 0.5$, nonEU(r) =

$W(\boldsymbol{D})U(1-(1-r)^2) + (1-W(\boldsymbol{D}))U(1-r^2)$.
or
$w(B(\boldsymbol{D}))U(1-(1-r)^2) + (1-w(B(\boldsymbol{D})))U(1-r^2)$.

Can always write $B(\boldsymbol{D}) = w^{-1}(W(\boldsymbol{D}))$,
so $W(\boldsymbol{D}) = w(B(\boldsymbol{D}))$.)

Is '92 prospect theory, = Schmeidler (‘89).
Includes multiple priors (Wald ’50; Gilboa & Schmeidler ’89);
For binary gambles: Einhorn & Hogarth ’85; Pfanzagl ’59; Luce (’00 Chapter 3); Ghirardato & Marinacci (’01, "biseparable").

**Theorem.** Under nonexpected utility with ambiguity,

$$r_{\boldsymbol{D}} = \frac{w(B(\boldsymbol{D}))}{w(B(\boldsymbol{D})) + (1 - w(B(\boldsymbol{D})))\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}$$

**Reversed (explicit) expression:**

$$B(\boldsymbol{D}) = w^{-1}\left(\frac{r}{r + (1-r)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\right)$$

# How bet on ***D*** now? [Ambiguity, nonEUA].

$r^{EV} = 0.75$.

$r^{EU} = 0.69$.

$r^{nonEU} = 0.61$.

Similarly,

$r^{nonEUA} = 0.52$ (under plausible assumptions).

r's are close to insensitive fifty-fifty.

"Belief" component B(***D***) = $w^{-1}(W) = 0.62$.

go to p. 15,
with figure of
R(p)

B(***D***): ambiguity attitude $\supset/=/\neq$ beliefs??
Before entering that debate:
How measure B(***D***)?
Our contribution: through proper scoring rules with "risk correction."

This ends Part I.

# Part II. Deriving Theoretical Concepts from Empirical Observations of r

We reconsider reversed (explicit) expressions:

$$p = w^{-1}\left(\frac{r}{r + (1-r)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\right)$$

$$B(\boldsymbol{D}) = w^{-1}\left(\frac{r}{r + (1-r)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\right)$$

Corollary. $p = B(\boldsymbol{D})$ if related to the same r!!

Our proposal takes the best of several worlds!

Need not measure U,W, and w.

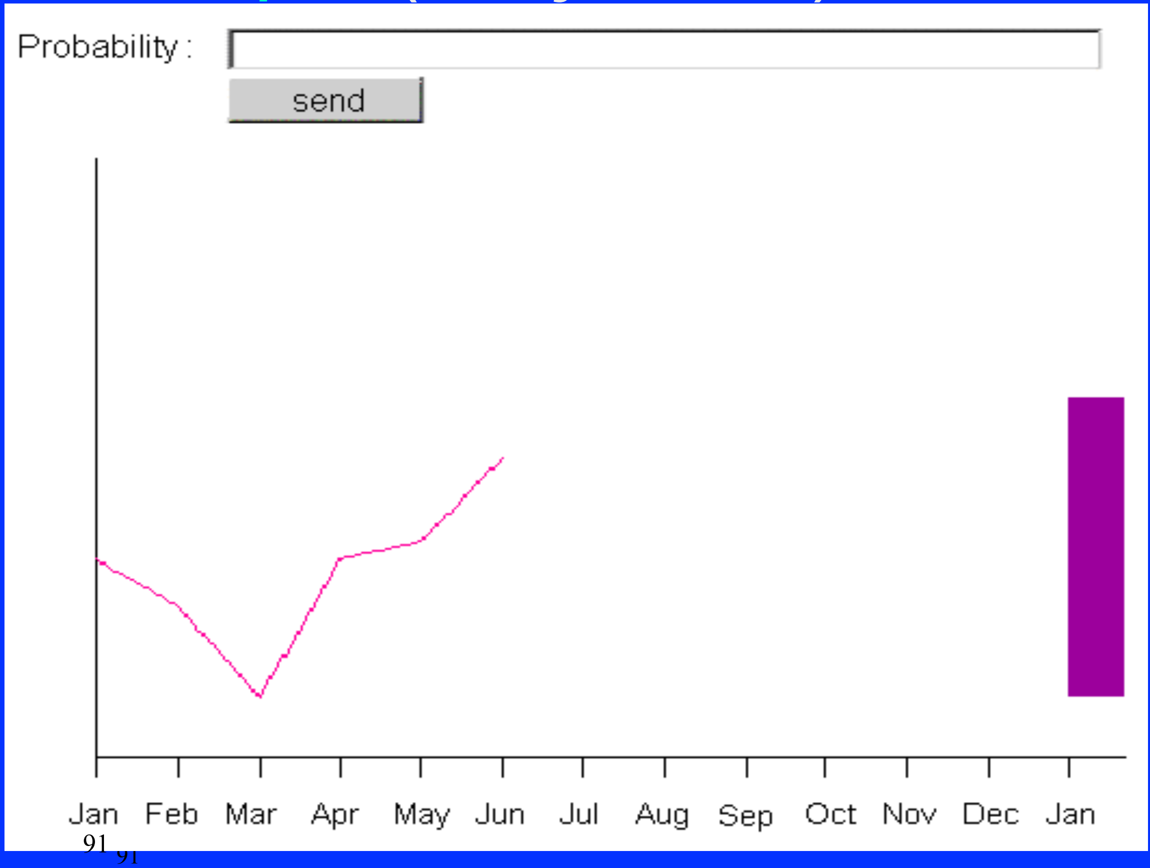
Get "matching probability" without measuring indifferences (BDM …; Holt 2006).

Calibration without needing many repeated observations.

Get ambiguity attitude without measuring U,w.

Do all that with no more than simple proper-scoring-rule questions.

# Example (subject 25)




stock 20, CSM certificates dealing in sugar and bakery-ingredients. Reported probability:
r = 0.75

For objective probability p=0.70, subject 25 also reported probability r = 0.75.
Conclusion: B(elief) of ending in bar is 0.70!
We simply measure the R(p) curves, and use their inverses: is risk correction.

Directly implementable empirically. We did so in an experiment, and found plausible results.

# Part III. Experimental Test of Our Correction Method

# Method

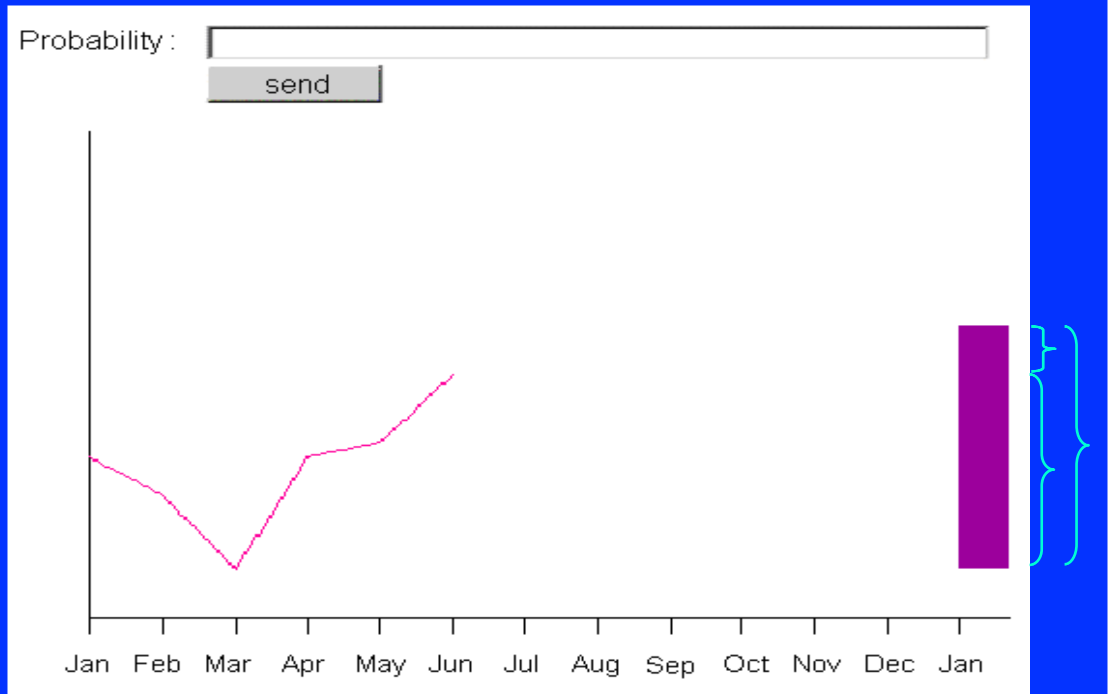
Subjects. N = 93 students.

Procedure. Computarized in lab.
Groups of 15/16 each.
4 practice questions.

# Stimuli 1. First we did proper scoring rule for unknown probabilities. 72 in total.

| Probability | Your score if statement is true | Your score if statement is not true |
|---|---|---|
| 27% | 4671 | 9271 |
| 28% | 4816 | 9216 |
| 29% | 4959 | 9159 |
| 30% | 5100 | 9100 |
| 31% | 5239 | 9039 |
| 32% | 5376 | 8976 |
| 33% | 5511 | 8911 |
| 34% | 5644 | 8844 |
| 35% | 5775 | 8775 |
| 36% | 5904 | 8704 |

send




For each stock two small intervals, and, third, their union. Thus, we test for additivity.

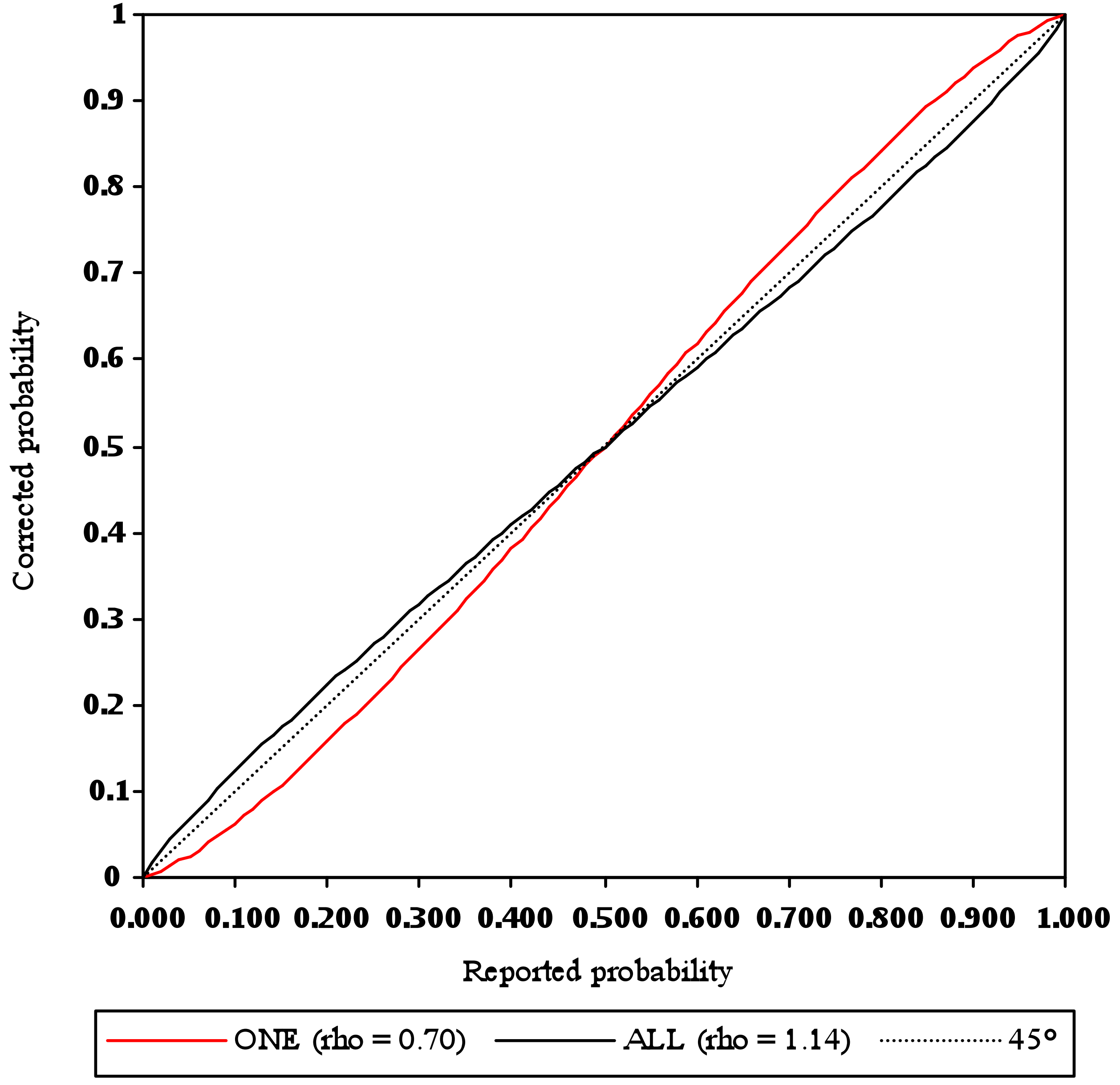
Stimuli 2. Known probabilities:
Two 10-sided dies thrown.
Yield random nr. between 01 and 100.
Event ***D***: nr. ≤ 75 (p = 3/4 = 15/20) (etc.).
Done for all probabilities j/20.

Motivating subjects. Real incentives.
Two treatments.

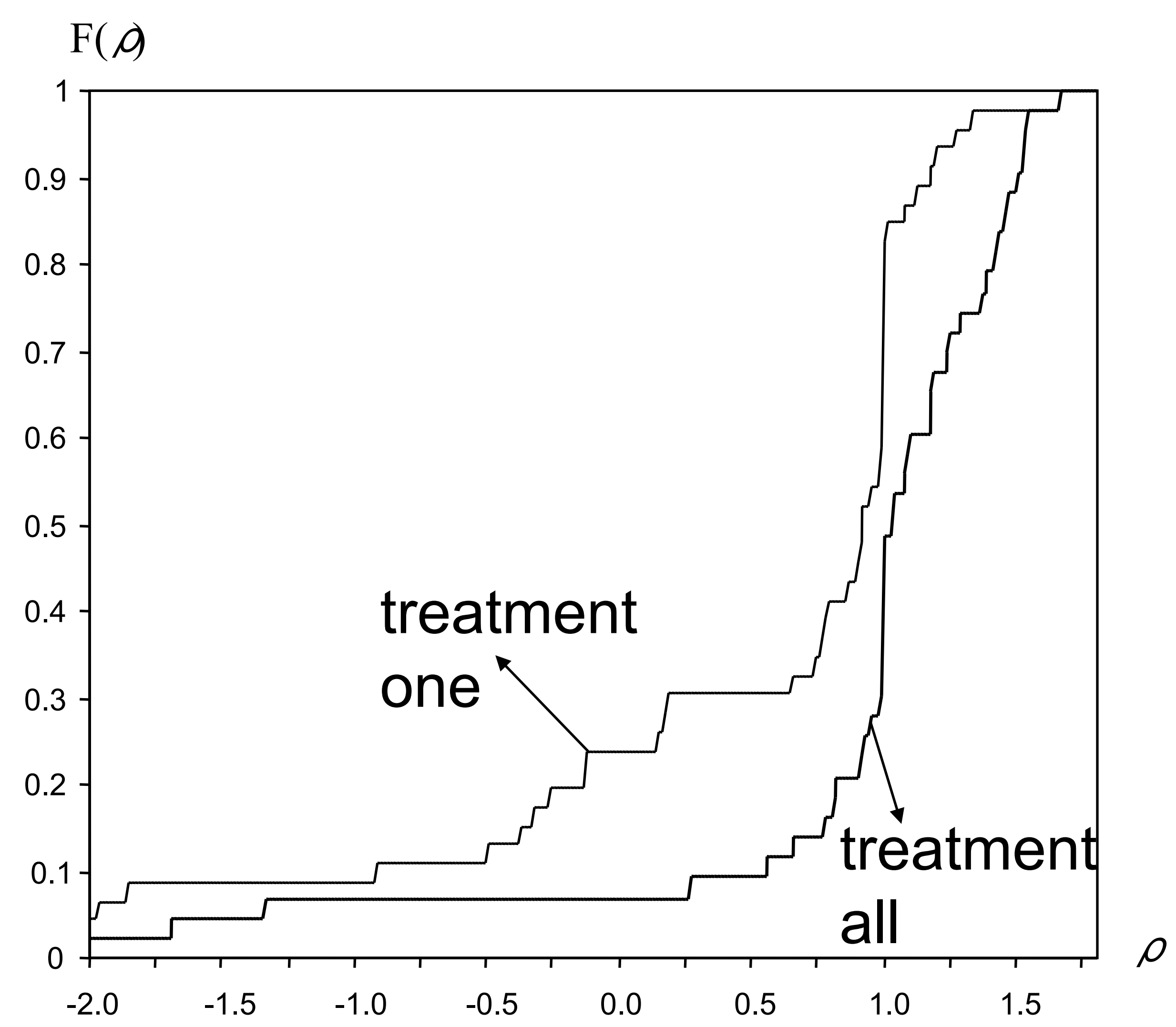
1. All-pay. Points paid for all questions.
   6 points = €1.
   Average earning €15.05.
2. One-pay (random-lottery system).
   One question, randomly selected afterwards, played for real. 1 point = €20. Average earning: €15.30.

# Results

(of group average; at individual level more corrections)

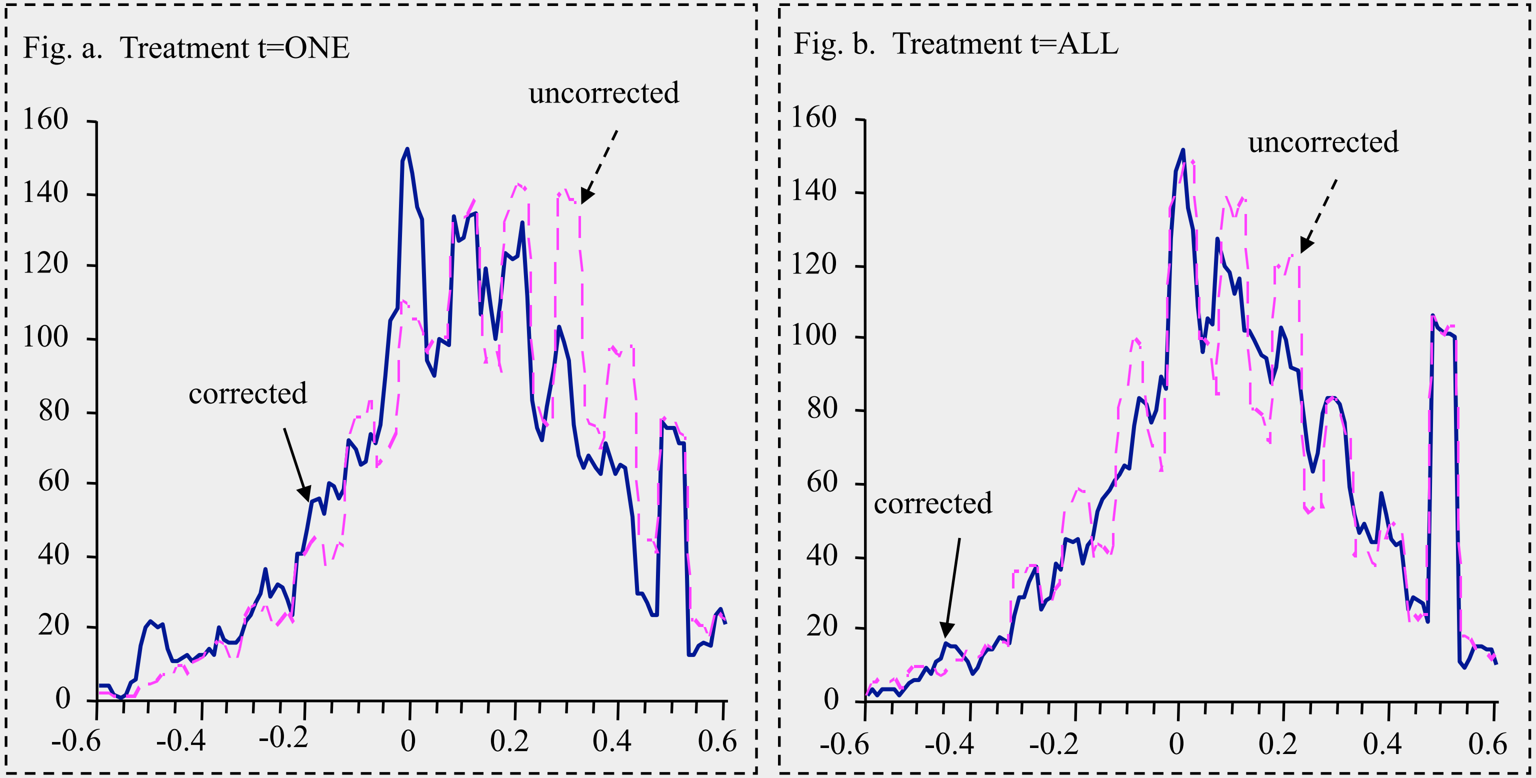



# Average correction curves




# Individual corrections

Figure 9.1. Empirical density of additivity bias for the two treatments




For each interval [(j-2.5)/100, (j+2.5)/100] of length 0.05 around j/100, we counted the number of additivity biases in the interval, aggregated over 32 stocks and 89 individuals, for both treatments. With risk-correction, there were 65 additivity biases between 0.375 and 0.425 in the treatment t=ONE, and without risk-correction there were 95 such; etc.

Corrections reduce nonadditivity, but more than half remains: ambiguity generates more deviation from additivity than risk.

Fewer corrections for Treatment t=ALL. Better use that if no correction possible.

## Summary and Conclusion

- Modern risk&ambiguity theories: traditional proper scoring rules are heavily biased.
- We correct for those biases. Benefits for proper-scoring rule community and for risk- and ambiguity theories.
- Experiment: correction improves quality; reduces deviations from ("rational"?) Bayesian beliefs.
- Do not remove all deviations from Bayesian beliefs. Beliefs are genuinely nonadditive/ nonBayesian/sensitive-to-ambiguity.

The end.