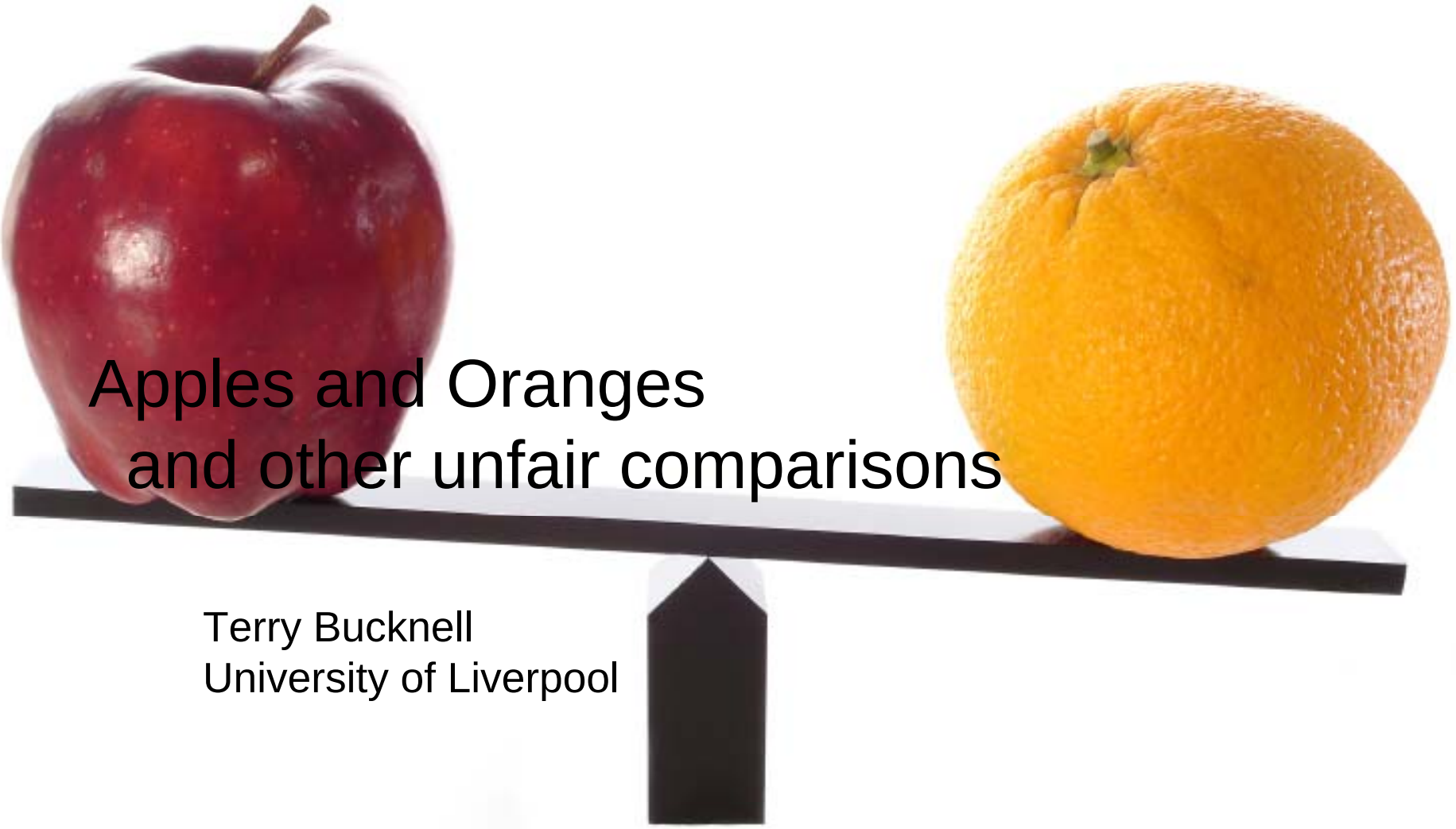


Apples and Oranges and other unfair comparisons

Terry Bucknell
University of Liverpool

Or... Garbage In, Gospel Out

Garbage In, Gospel Out is a more recent expansion of the acronym [GIGO – Garbage In, Garbage Out]. It is a sardonic comment on the tendency to put excessive trust in “computerized” data, and on the propensity for individuals to blindly accept what the computer says. Because the data goes through the computer, people tend to believe it.

Decision-makers increasingly face computer-generated information and analyses that could be collected and analyzed in no other way. Precisely for that reason, going behind that output is out of the question, even if one has good cause to be suspicious. In short, the computer analysis becomes the gospel.

http://en.wikipedia.org/wiki/Garbage_In,_Garbage_Out

GIGO in action

An unswerving acceptance of cost per download:

- Less than £1 good; more than £1 bad
- Publisher: “what’s an acceptable cost per download?”
- Publishers setting targets to boost usage
- Publishers adding site features to boost usage

Sounds familiar?

- Impact Factor
  - A simple measure of citations, weighted by number of citable papers published
  - Author and publisher behaviour has changed to maximise IF
  - Now beloved of publishers' marketing departments
  - Regarded with cynicism / distrust by librarians
- Usage Factor
  - Publishers preparing for this by setting usage targets?

COUNTER usage statistics

- COUNTER aims to provide statistics that are:
  - Consistent – same definition of what / how to count
  - Credible – audited
  - Compatible – delivered in same formats
- This does *not* imply that COUNTER usage statistics from different platforms are *directly comparable*

Problems with cost per download

- Denominator issues – Usage
  - Platform effects, subject differences, type of content, amount of content, usage spikes
- Numerator issues – Cost
  - Currency fluctuations, new pricing models
- What does the cost per download formula *mean*?

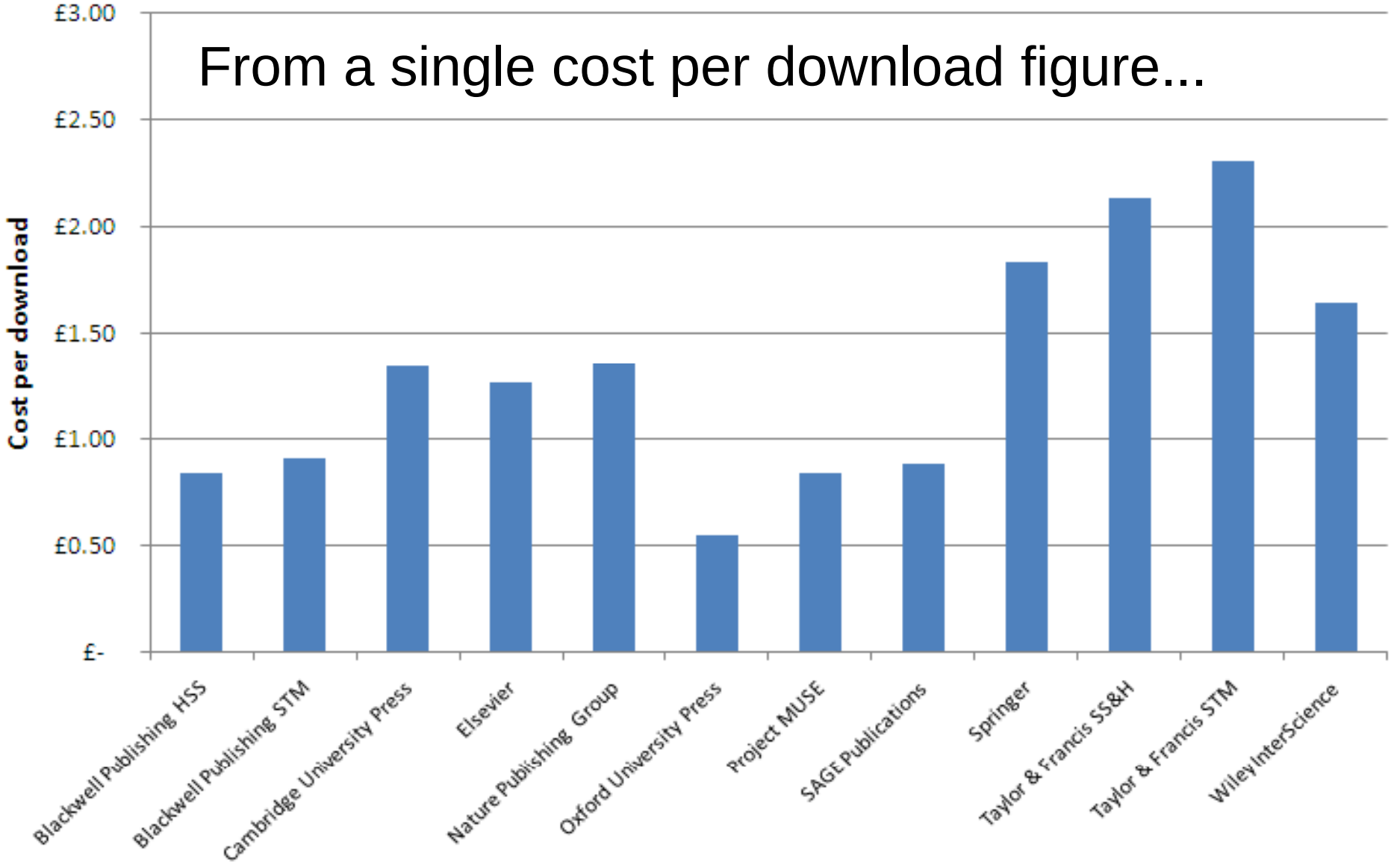
Usage issues – PDF/HTML ratio

- SpringerLink: 97% PDF, 3% HTML
- ScienceDirect: 59% PDF, 41% HTML
- nature.com: 65% PDF, 35% HTML
- Does landing page from Google/SFX/TOC lead to...
  - Abstract with links to HTML and PDF full-text, or
  - HTML full-text with link to PDF
- Are HTML & PDF downloads equal in *value*?

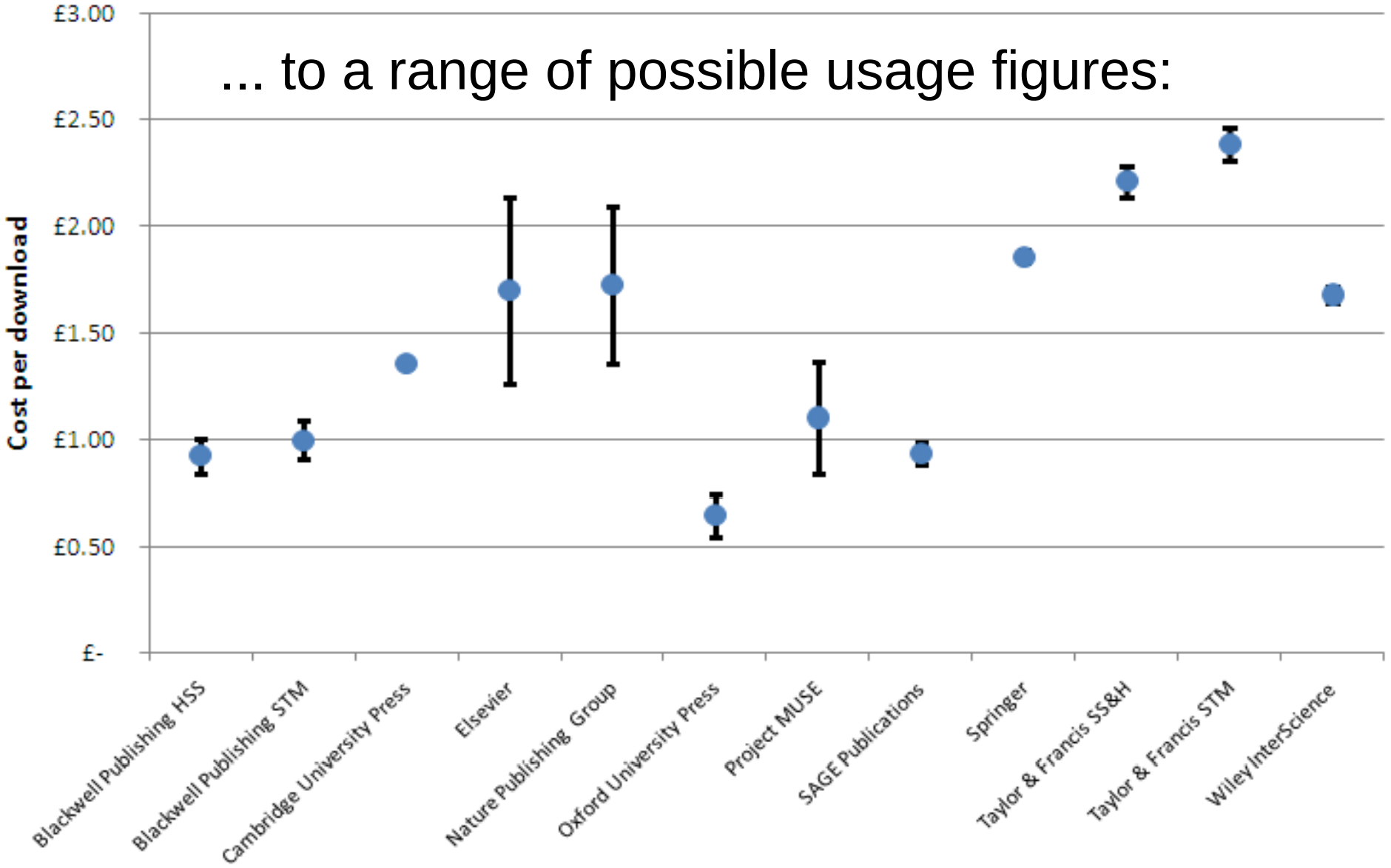
Usage issues – PDF/HTML ratio

- We cannot know how often an attempt to view an article leads to *both* an HTML view and a PDF view
- Quote *range* of usage, not just single value:
  - From: higher of *PDF Total* or *HTML Total*
  - To: *YTD Total*
- The closer the PDF:HTML ratio is to 50:50, the greater the uncertainty in the ‘true’ usage

From a single cost per download figure...



... to a range of possible usage figures:



Usage issues – subject bias

- *Tenopir & King* tell us that e.g. Medics read far more articles than Mathematicians, but read each article for less time
  - Usage stats measure only *number* of readings, not *duration*
- Do we only compare titles in the same subject area?
- Or are some subjects so bad that we should use document delivery / PPV rather than subscriptions?
- Selection by usage leads to an unbalanced collection...

If I was guided solely by usage statistics, I would cancel all my subscriptions to ~~humanities~~ maths journals, which ~~tend to publish far fewer issues per year than the monster science titles~~..... never seem to get read, no matter how much they cost

Terry Bucknell, Electronic Resources Manager at Liverpool University

Usage issues – spikes

- How do you know when a spike in usage is misuse and when it is genuine heavy use (e.g. a class)?
- Do you keep spikes or correct for them?
  - Keep for SCONUL stats?!
  - Correct for them when evaluating value?
  - Must record what you did for future comparisons!

Usage issues – aggregators

- Exclude, because aggregator usage is independent of your subscription?
- Include, because it reflects the total demand for a journal (risk of withdrawal from the aggregator)?
- Complicated by differing coverage between platforms

Cost issues – exchange rate variability

- Consider 3 journals in 2009:

	Cost	Downloads	cost per download
Journal A	£1,000	500	£ 2.00
Journal B	£1,000	500	£ 2.00
Journal C	£1,000	500	£ 2.00

- But A is *invoiced* in £, B in €, C in $

Cost issues – exchange rate variability

What if, in 2010...

- Each price is frozen
- Usage happens to be identical to 2009
- £ declines 20% against €
- £ gains 20% against $

Cost issues – exchange rate variability

- The result in 2010 is...

	Cost	Downloads	Cost per download
Journal A	£ 1,000	500	£ 2.00
Journal B	£ 1,250	500	£ 2.50
Journal C	£ 800	500	£ 1.60

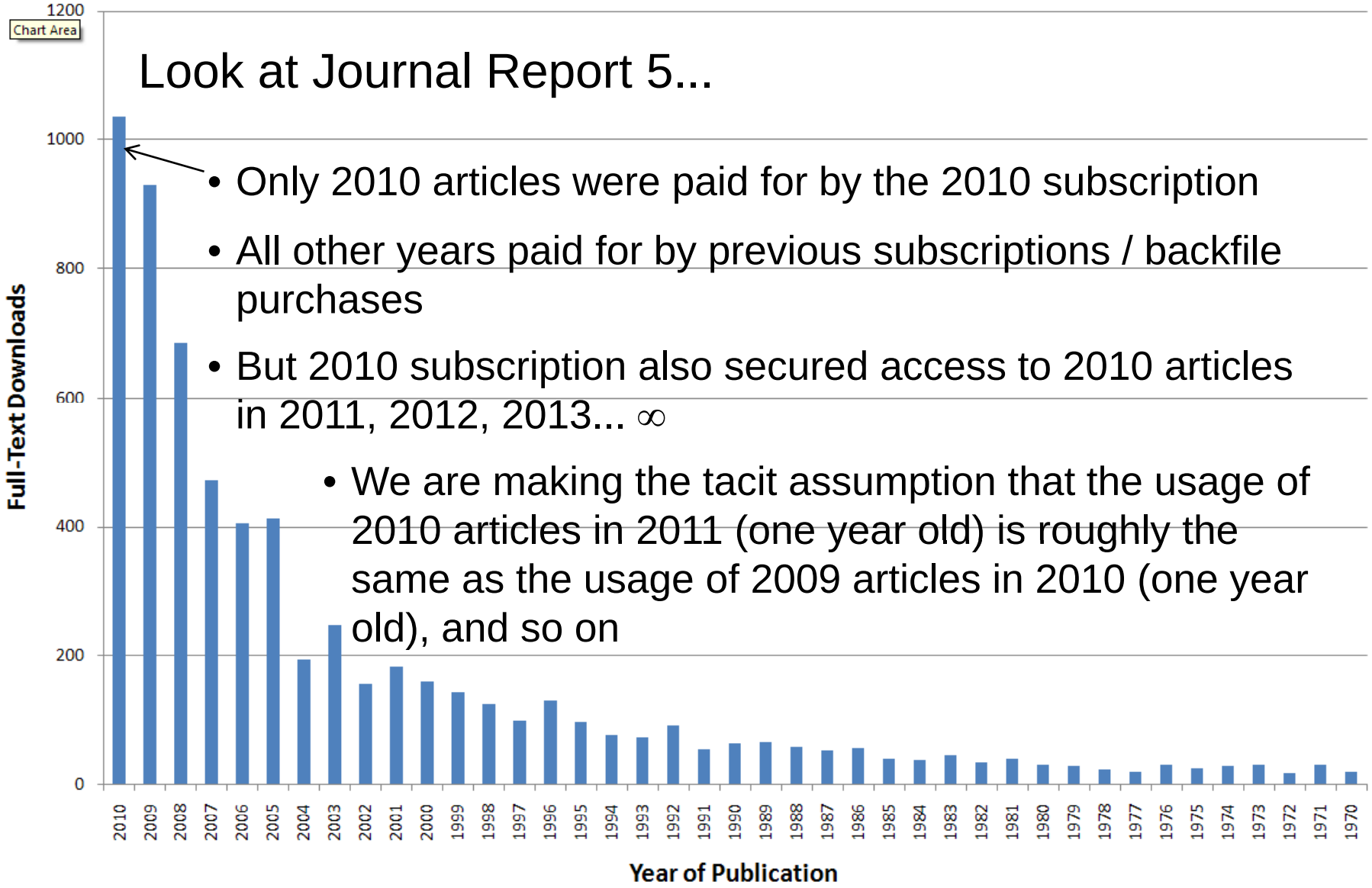
Cost issues – exchange rate variability

- B is now 25% more expensive than A
- C is now 20% cheaper than A
- B is 56.25% more expensive than C
- Do we hold publishers liable for currency markets?
- But it could all change back again next year!

Cost per Download (CPD)

- Just a crude attempt to normalise usage for differently-priced journals
- *Not* a rigorous unit cost calculation:
  - Cost of staffing support per journal: processing of any print, managing link resolver etc, monitoring usage, responding to enquiries
  - Only include usage paid for by that year's subscription: exclude OA articles, purchased backfiles, previous years' subscription content

Look at Journal Report 5...



- Only 2010 articles were paid for by the 2010 subscription
- All other years paid for by previous subscriptions / backfile purchases
- But 2010 subscription also secured access to 2010 articles in 2011, 2012, 2013... ∞
  - We are making the tacit assumption that the usage of 2010 articles in 2011 (one year old) is roughly the same as the usage of 2009 articles in 2010 (one year old), and so on

Implications of assumptions in CPD calculation

- Young journals see lower usage, higher CPD (intuitive but how do you correct for it?)
- For a *rigorous* CPD calculation should include all backfile usage (because it most closely approximates *all future* usage)
- For a *comparable* CPD calculation should always use same years of publication (but can't because few publishers provide JR5)

Decide what question you are asking

- “What is the unit cost of the articles downloaded as the result of this year’s subscription?”
  - A rigorous CPD calculation, using all past usage as an approximation to all future usage, corrections for OA articles, etc.
- “Could I have got away without subscribing this year?”
  - Just include usage for the years of publication that you didn’t already have access to (if you have a JR5 report)

Key Messages

- Cost per download is unavoidably imprecise
- Try to give an indication of the uncertainty
- Provide a commentary alongside the stats
- Use usage stats as a first filter
- Other factors need to be considered too

The Approximate Time mobile app:

Not:

12:41:38

But...

The Approximate Cost per Download app:

Not:

Journal A: £1.65
Journal B: £1.37

But...

Both journals around the £1.50 mark

Thank you for listening