ONLINE APPENDIX

1. The type of articles included in the overview of reviews for both WTP-studies and surrogacy literature (Table A1)
2. Search terms used when identifying papers for review.
3. Benefit transfer analyses.
4. Currency adjustments for WTP estimates.

Table A1. The type of articles included in the overview of reviews for both WTP-studies and surrogacy literature.

Topic of review	Type of literature review	Reference
WTP	meta-analysis	Amuakwa-Mensah et al. 2018
WTP	meta-analysis	Richardson & Loomis 2009
WTP	meta-analysis	Martin-Lopez et al. 2008
WTP	meta-analysis	Lindhjem & Tuan 2012
WTP	meta-analysis	Lindhjem & Tuan 2012
WTP	meta-analysis	Jacobsen & Hanley 2009
WTP	meta-analysis	Hjerpe et al. 2015
WTP	meta-analysis	Lindhjem 2006
WTP	meta-analysis	Barrio & Loureiro 2010
WTP	meta-analysis	Ojea E., Loureiro M.L. 2011
WTP	meta-analysis	Žáková Kroupová et al. 2016
WTP	meta-analysis	Subroy et al. 2019
WTP	meta-analysis	Nijkamp et al. 2008
surrogacy	meta-analysis	Eglington et al. 2012
surrogacy	meta-analysis	Wolters et al. 2006
surrogacy	meta-analysis	Westgate et al. 2014
surrogacy	meta-analysis	Westgate et al. 2017
surrogacy	meta-analysis	de Morais et al. 2018
surrogacy	meta-analysis	Castagneyrol and Jactel 2012
surrogacy	meta-analysis	Velghe and Gregory- Eaves 2013
surrogacy	essay	Heino 2010
surrogacy	systematic review	Rodrigues and Brooks 2007
surrogacy	meta-analysis	Mellin et al. 2011
surrogacy	meta-analysis	Schmit et al. 2005
surrogacy	systematic review	Gao et al. 2015
surrogacy	essay	Halme et al. 2017
surrogacy	essay	Carignan and Villard 2002
surrogacy	essay	Roberge and Angelstam 2004
surrogacy	meta-analysis	Branton and Richardson, 2011
surrogacy	systematic review	MacPherson et al. 2018
surrogacy	essay	Noss et al. 1996
surrogacy	essay	Dalerum et al. 2008
surrogacy	essay	Linnell et al. 2000
surrogacy	essay	Sergio et al. 2008
surrogacy	meta-analysis	Root-Bernstein and Ebensperger 2013
surrogacy	essay	Stoner et al. 2007
surrogacy	meta-analysis	Prevedello et al. 2018
surrogacy	systematic review	van der Hoek et al. 2019
surrogacy	essay	Peres 2000
surrogacy	systematic review	Favreau et al. 2006
surrogacy	essay	Zacharias and Roff 2001
surrogacy	meta-analysis	Stringer and Gaywood, 2016
surrogacy	essay	Rosell et al., 2005
surrogacy	essay	Janiszewski et al., 2014
surrogacy	essay	Davic and Welsh, 2004
surrogacy	meta-analysis	Arredondo-Núñez et al., 2009
surrogacy	essay	Kotliar et al., 1999

2. Search terms used when identifying papers for review

WTP meta-review:

TS=(("willingness to pay" OR "contingent valuation" OR "non-market valuation" OR"stated preference" OR "choice experiment*") AND (conservation OR species OR ecosystem* OR biodiversity OR "biological diversity") AND (review* OR meta-analys*)) Scopus = 261, WoS = 256, EBSCO = 150 (with duplicates removed), combined = 384

Surrogacy meta-review:

meta-analysis or systematic review or literature review) AND "surrogate species"
EBSCO=5, Scopus=4, WOS=7

meta-analysis" OR "systematic review" OR literature review) AND "umbrella species"
EBSCO=8, Scopus = 3, WOS=11

meta-analysis" OR "systematic review" OR literature review) AND "keystone species"
EBSCO=36, Scopus = 11, WOS=53

meta-analysis" OR "systematic review" OR literature review) AND "indicator species"
EBSCO=121, Scopus = 10, WOS=49

The papers were searched from EBSCO, Scopus and Web of Science as well as using Google Scholar. Some of the papers were identified from the reference list of other reviews or meta-analyses.

3. Benefit transfer analyses

3.1 Chosen models

In general, when multiple models were presented in a study, we chose the one with the highest R^2. If there were differences in significance of relevant variables, we preferred the model that allowed predicting values for multiple flagship types. For example, in Ojea & Loureiro (2015) the model that was best from the perspective of scope sensitivity (the sensitivity of the model to the quantity of the good valued, i.e., does WTP depend on how much is protected), which was the topic of their study, did not have significant coefficients for the variables representing flagship types, so we chose a model that was scope-insensitive but where "forest" and "species" had a significant effect. Additionally, if multiple models had been compared, we preferred models with logarithmic scales and that considered the panel structure of the data (multiple observations coming from the same publication or author).

3.2 Common variable values

Species were either all threatened/endangered/rare to begin with (Amuakwa-Mensah et al. 2018, Richardson & Loomis 2009, Tuan & Lindhjem 2008), or the IUCN category was not significant (Martin-Lopez et al. 2008). Only one study had a significant coefficient for 'threatened species', and we produced BT estimates for both types (Subroy et al. 2019). We chose 'contingent valuation' as the valuation method, which was an option shared by five studies. We chose 'tax' as the payment vehicle, because it was the only shared alternative across studies in which it was significant, even though voluntary payments would be more relevant for studying donation behavior. We used 'dichotomous choice' as the elicitation technique as it was most commonly included. Payments per household were more common, and we used it as our choice when there was a variable for payment per household or per person. It also varied whether payments were lump sums, monthly, yearly for a determined or undetermined time, and we used yearly payments (or 'other when the option was 'lump sum' vs. 'other'), with the exception of Subroy et al. (2019) who studied lump sum payments only. We also used a common year, 2006, that was within the range of study years in all MAs, for all studies where year was a significant covariate. For four studies we set 'resident'=1 as opposed to visitor status, to avoid overestimates produced by ecotourism-focused studies (Richardson and Loomis, 2009, Amuakwa-Mensah et al. 2018, Subroy et al., 2019, Žáková Kroupova et al., 2016).

3.3 Variable values for unique or rare variables

We mostly used sample means for methodological variables unique to specific studies whenever they were available, which is common practice in BT and has been found to increase the transfer error only slightly (Stapler & Johnston 2009). For example, Lindhjem & Tuan (2012) found negative coefficients in their models for the variable 'nonparametric methods', and we used the sample mean of 0.07 of the dataset in the BT estimates.

GDP per capita was significant in three studies, and we used the sample mean from Hjerpe et al. (34 614 USD2010) also for Jacobsen & Hanley (2009), Barrio & Loureiro (2010) as well as Ojea & Loureiro (2011). In Lindhjem (2006) we used MSc thesis=0, Unpublished=0, because other studies used almost exclusively peer-reviewed primary papers and typically did not have variables for these, and the variable Urban=0 because our focus, as well as those of most MAs, was not on urban green spaces. Lindhjem (2006) also had a number of other significant variables, for which we used sample means (Table 1 in manuscript). In the case of Martin-Lopez et al. (2008) we looked up high and low values for the variable "eye size" as mean +/- 1 SD from the original article the data were extracted from (Howland et al. 2004). For Lindhjem & Tuan (2012) the dummy for Southeast Asia had a negative coefficient in the model for habitats, and we set the dummy to zero to obtain estimates closer to other studies focusing on wealthier parts of the world. Their model for species had a negative

coefficient for Australia, and the authors hypothesized this was due to species with lower charisma in Australian studies. Therefore we set this dummy to zero as we are more interested in charismatic flagship species. For other geographical variables we used sample means. Barrio & Loureiro (2010) had a positive coefficient for a recreation services variable, which we set to zero, and a negative coefficient for forest land area in the country, where we used the sample mean of 83 700 ha. Zakova Kroupova et al. (2016) included several primary studies where WTP was asked as e.g. NP visitor fees, but this was accounted for by setting the (significant) dummy variable to 'resident' in the BT calculation. In the same study the variable for the share of agriculture in gross value added for the country had a positive significant impact, but sample mean was not given. Instead we used a mean between two values that were given as examples of limits below and above which the WTP values were "lower" or "higher": 1% and 3.3%, making a mean of 2.15%. Following the same logic, we used 15% as the mean between 10% and 20% given as examples of low and high shares of LFA (Less-Favoured Areas) subsidies.

3.4 Other details

The authors of meta-analyses had typically used significance $p<0.1$ for their models, and we followed it as well when deriving BT estimates based on models that contained non-significant covariates.

4. Currency adjustments for WTP estimates

We used various resources for adjusting currencies from different years to 2019 prices (https://www.ssb.no/en/kpi, https://www.ekonomifakta.se/, https://www.bankofengland.co.uk/monetary-policy/inflation/inflation-calculator, https://www.inflationtool.com/, https://www.usinflationcalculator.com/). Because majority of the papers reported their results in USD, other currencies were then converted to USD using PPP adjustment for 2019 (https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm). We then converted all the results in USD into Euros using the PPP-adjusted exchange rate for 2019, 1 USD = 0.684 EUR). Results originally given in Euros were corrected only for inflation.