Choice of choice models: Theory of signal detectability outperforms Bradley-Terry-Luce choice model

Diana E Kornbrot Corresp., [1] , **George J Georgiou** [1] , **Mike Page** [1]

[1] Psychology, University of Hertfordshire, Hatfield, UK

Corresponding Author: Diana E Kornbrot
Email address: d.e.kornbrot@herts.ac.uk

Identifying the best framework for two-choice decision-making has been a goal of psychology theory for many decades (Bohil, Szalma, & Hancock, 2015; Macmillan & Creelman, 1991). There are two main candidates: the theory of signal detectability (TSD) (Swets, Tanner Jr, & Birdsall, 1961; Thurstone, 1927) based on a normal distribution/probit function, and the choice-model theory (Link, 1975; Luce, 1959) that uses the logistic distribution/logit function. A probit link function, and hence TSD, was shown to have a better Bayesian Goodness of Fit than the logit function for every one of eighteen diverse psychology data sets (Open-Science-Collaboration, 2015a), conclusions having been obtained using Generalized Linear Mixed Models (Lindstrom & Bates, 1990; Nelder & Wedderburn, 1972) . These findings are important, not only for the psychology of perceptual, cognitive and social decision-making, but for any science that use binary proportions to measure effectiveness, as well as the meta-analysis of such studies.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26978v1 | CC BY 4.0 Open Access | rec: 5 Jun 2018, publ: 5 Jun 2018

Choice of Choice Models: Theory of Signal Detectability Outperforms Bradley-Terry-Luce Choice Model

Diana E. Kornbrot, George J Georgiou, Mike Page

University of Hertfordshire, UK

Corresponding author

Diana Kornbrot

d.e.kornbrot@herts.ac.uk

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26978v1 | CC BY 4.0 Open Access | rec: 5 Jun 2018, publ: 5 Jun 2018

Abstract

Identifying the best framework for two-choice decision-making has been a goal of psychology theory for many decades (Bohil, Szalma, & Hancock, 2015; Macmillan & Creelman, 1991). There are two main candidates: the theory of signal detectability (TSD) (Swets, Tanner Jr, & Birdsall, 1961; Thurstone, 1927) based on a normal distribution/probit function, and the choice-model theory (Link, 1975; Luce, 1959) that uses the logistic distribution/logit function. A probit link function, and hence TSD, was shown to have a better Bayesian Goodness of Fit than the logit function for every one of eighteen diverse psychology data sets (Open-Science-Collaboration, 2015a), conclusions having been obtained using Generalized Linear Mixed Models (Lindstrom & Bates, 1990; Nelder & Wedderburn, 1972). These findings are important, not only for the psychology of perceptual, cognitive and social decision-making, but also for any science that use binary proportions to measure effectiveness, as well as the meta-analysis of such studies.

Choice of Choice Models: Theory of Signal Detectability Outperforms Bradley-Terry-Luce Choice Model

The aim of this report is to compare two different theoretical frameworks for modelling decision-making: namely, those based on the theory of signal detection (TSD) (Swets, et al., 1961; Thurstone, 1927) and hence the normal distribution alongside its associated quantile function called the probit function; and those based on Bradley-Terry-Luce Choice models (BTL) (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Link, 1975) and hence the associated logit function. The comparison is made using 18 data sets that use binary proportions as a response (Open-Science-Collaboration, 2015a, 2015b). Bayesian Information Criterion (BIC) Goodness-of-Fit measures, from Generalized Linear Mixed Models (GLMMs), were used to make the comparisons.

TSD and BTL have long been theoretical rivals. It had been thought that they were almost indistinguishable empirically for two-choice tasks because, except at their extremes, probit and logit functions are sufficiently similar that no reliable empirical discrimination of the two has previously been found, e.g. (Bohil, et al., 2015; Macmillan & Creelman, 1991). Although work with ordinal categorical judgements does suggest a superiority of TSD as early as 1978 (Kornbrot, 1978), since then, much theoretical work has concentrated on development of either the signal detection (Killeen, Taylor, & Treviño, 2018) or the choice framework (Bohil, et al., 2015). In spite of their similarities, the generic mechanisms that lead to logit and probit distributions are different.

One kind of mechanism that leads to a logit distribution is the random walk with a drift rate applied according to the evidence available (Wald, 1947): This has been frequently proposed for perceptual discrimination (Laming, 1968; Luce, 1986). It is known that ‘pure’ random walks are not sufficient as they predict that if barrier locations are held constant there will be identical distributions for a specific response, whether given correctly or in error. Several modifications have been suggested to account for the finding that this almost never happens. Specific examples include error-correcting models, where people move their barrier location after an error. We do not know if such mechanisms would produce results more compatible with a probit function.

Another theoretical mechanism that generates choice data compatible with a logit function comes from research on category judgments, such as 'word' or 'non-word' in lexical decision, 'cheat' or 'honest' in social decisions, or 'old' or 'new' in memory studies. One kind of memory model suggests that, with experience of members of each of two categories, observers build up a set of exemplars for each category. They then compare any new test exemplar with the memorized exemplars to establish the best-matching memorized exemplar, giving the test exemplar the category label associated with that best-matching exemplar. Choice behaviour can then be expected to be well modelled by a logit function under specific conditions. These include when experience contains multiple (i.e., repeated) presentations of the same training exemplars and where the degree-of-match of only the best matching exemplar of any given category is considered in the choice process. Technically, this is because the *extreme* value (e.g., highest value) across a number of variables that are identically normally distributed, is characterized by the Gumbel distribution. The difference between two Gumbel distributions is a logistic distribution, for which the logit, and not the probit, is the appropriate distribution function (Page, 2000). By contrast, the *pooled* distribution of a set of variables that are normally distributed is, of course, normally distributed itself, and the probit function is the appropriate quantile function. A finding that the probit function provides for better-fitting models would suggest, therefore, that in classification tasks, match-information across many learned exemplars (particularly where individual exemplars are repeatedly presented during learning) is more likely to be pooled, as opposed to being reduced just to the best-matching exemplar from each category.

Pooling over relevant mental representations is just one of many mechanisms that might generate a normal distribution of, say, match values to a given test stimulus. This is because the central limit theorem suggests that the normal distribution occurs whenever multiple sources of variable information contribute to some feature. In the classical signal-detectability account of a perceptual experiment, the representation in the human brain of a sequence of physically identical stimuli has a normal distribution, hence d' is the discrimination measure of choice. This signal detection model can be generalized to any classification task (Ratcliff, 1978; Ratcliff & McKoon, 2008). A finding of probit superiority would, therefore, generally support models that have multiple sources of information or 'activation' even for both simple perceptual and complex cognitive tasks. Current paradigms do not enable us to distinguish whether these multiple sources operate in the primary representation of stimuli or in the criteria setting that is an integral and unavoidable part of any of these tasks.

In any event, a method that can reliably distinguish these frameworks has considerable theoretical importance. Since there are persuasive arguments for both logit and probit as the appropriate function to apply when assessing evidence in choice tasks, we had no predictions as to which framework would 'win'. Indeed, we considered it quite possible that the best model would depend on the task, maybe TSD probit for more perceptual tasks, and BTL logit for more cognitive tasks.

Method

Data sets

The data-sets were downloaded from the Open Science Collaboration website (Open-Science-Collaboration, 2015a, 2015b). There were 100 data-sets. We chose all those (18) that met the criteria that the response variable was effectively a proportion (i.e., the number of trials meeting some specified criterion from a fixed number of opportunities), and that the published analysis was ANOVA. This was because one of our goals was comparing the descriptive and inferential results of ANOVA and GLMM analyses. The topics covered a range of social and cognitive areas of psychology, and used several designs with between 1 and 4 factorial predictors, some varying between group and some repeated over participants. Details of the ANOVA/GLMM comparison are available in a separate manuscript. Table 1 summarizes properties of studies.

____

Insert Table 1 about here

Analysis Methods

GLMM analyses were conducted on each data-set using the SPSS procedure MIXED. Each analysis was run twice: once with a probit link function, once with logit. BIC goodness-of-fit measures were compared.

Results

Table 2 shows the design and resulting BIC values and the ratio of logit BIC to probit BIC. The probit link gave the best fit, that is, it had the lowest BIC, for all studies. The ratio varied from 1.1 to 16.1 for the 17 cases with positive BIC. A negative BIC was obtained for study-15 probit, as is possible with these kinds of model, so probit was best for this study also.

____

Insert Table 2 about here

Discussion

The signal-detection framework is shown to be superior to the choice framework across all 18 data sets. This is a serendipitous finding. Our initial aim was to show that *any* binomial GLMM was better than standard ANOVA, as was indeed the case. As noted earlier, we had no a priori prediction between logit and probit, so the unequivocal favouring of the normal distribution, as instantiated by the probit link, came as a big surprise. This finding has implications both for theories of psychological discrimination and for methods of choosing between rival theories in any science.

For psychological discrimination the TSD framework is favoured across a wide range of diverse Tasks (Table 1). TSD supports models that have multiple sources of information or ‘activation’ even for the simplest tasks. More specifically, for classification tasks it suggests that match-information across many learned exemplars is more likely to be pooled, rather than being reduced just to the best-matching exemplar from each category.

The psychological discrimination problem is structurally very similar to medical meta-analysis problems where the response variable is binary (e.g., dead or alive, disease progressed or not, etc.). Many, if not most, such meta-analyses use *log (odds ratios)* which are logits, although some do use probits. We have not been able to discern how the strategic choice between *log (odds ratio)* and probit is made. Our findings suggest a comparison between logit and probit should be a routine part of meta-analyses of binary proportions. This should ensure that the best-fitting model is used to draw conclusions about potentially life-threatening conclusions.

These results also show the benefit of using GLMMs to identify best models for proportion data in *any* area of science, including meta-analysis. This is a considerable methodological advance and a very practical reason for using GLMMs.

Conclusions

We draw the following conclusions.

The signal-detection framework is superior to the choice framework for modelling of proportions as a response variable across a wide range of psychological domains.

Generalized Linear Mixed Models constitute a method of analysis with statistical theoretical support, which (while not new) deserves to be used more widely by psychological and other sciences.

The results suggest that probit links may be more useful for meta-analysis than the more prevalent log-likelihood methods that use logit links. In any event, meta-analyses should compare logit and probit links for goodness-of-fit. We were unable to find any meta-analyses where such a comparison occurred.

These results contribute to resolving a major issue in psychology, and suggest a powerful method of identifying best models in science generally.

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700-765. https://doi.org/10.1037/0033-295X.113.4.700

Bohil, C. J., Szalma, J. L., & Hancock, P. A. (2015). Psychophysical Methods and Signal Detection: Recent Advances in Theory. In J. L. Szalma, M. W. Scerbo, P. A. Hancock, R. Parasuraman & R. R. Hoffman (Eds.), *The Cambridge Handbook of Applied Perception Research* (pp. 22-38). Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-applied-perception-research/psychophysical-methods-and-signal-detection-recent-advances-in-theory/D7D5122A8031F0529D3E73D4D4F5B4C4

Killeen, P. R., Taylor, T. J., & Treviño, M. (2018). Subjects adjust criterion on errors in perceptual decision tasks. *Psychological Review, 125*(1), 117-130.

Kornbrot, D. E. (1978). Theoretical and Empirical Comparison of Luces Choice Model and Logistic Thurstone Model of Categorical Judgment. [Article]. *Perception & Psychophysics, 24*(3), 193-208. https://doi.org/10.3758/BF03206089

Laming, D. R. J. (1968). *Information theory of choice reaction times*. London: Academic Press.

Lindstrom, M. J., & Bates, D. M. (1990). NONLINEAR MIXED EFFECTS MODELS FOR REPEATED MEASURES DATA. *Biometrics, 46*(3), 673-688. <Go to ISI>://BCI:BCI199090123598

Link, S. W. (1975). The Relative judgment theory of two-choice reaction time. *Journal of Mathematical Psychology, 12*, 114-135.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1986). *Response times*. Oxford: Clarendon Press.

Macmillan, N. A., & Creelman, D. C. (1991). *Detection theory: a user's guide*. Cambridge: Cambridge University Press.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society A, 135*, 370-384.

Open-Science-Collaboration. (2015a). Estimating the reproducibility of psychological science. *Science, 349*(6251). http://www.sciencemag.org/content/349/6251/aac4716.abstract

Open-Science-Collaboration. (2015b, 2015-08-28 00:00:00). Estimating the reproducibility of psychological science. DATA. *Science* Retrieved 2-Oct-2017, 2017, from https://osf.io/ezcuj/wiki/Replicated Studies/

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23*(4), 443-467. https://www.cambridge.org/core/article/connectionist-modelling-in-psychology-a-localist-manifesto/65F9E3CEC90E0C80A46B25E0028BCFE3

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 238-255.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation, 20*(4), 873-922. <Go to ISI>://MEDLINE:18085991

Swets, J. A., Tanner Jr, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*(5), 301-340.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Wald, A. (1947). *Sequential Analysis.* New York: Wiley.

Supplementary Material

EXCEL workbook **ProportionRawAll.xlsx** contains raw data with one sheet for each study, specifying the study number, *study*; participant number, *pno*; all between predictors, *b1, b2* etc.; all within predictors, *w1, w2*, etc.; the number of observations meeting the criterion, *freq*, and the number of opportunities *Nmax*.

Table 1(on next page)

Properties of data sets

Id numbers, authors, URLs and main topics/themes for 18 data sets.

Table 1

Properties of data sets

ID	First Author	Project URL	Topic
002	Morris	HTTPS://OSF.IO/RMVK5/	Repetition blindness for nonwords
003	Liefooghe	HTTPS://OSF.IO/4DVZB/	Working memory costs of task switching
004	Storm	HTTPS://OSF.IO/8J9CG/	Fast relearning, retrieval-induced forgetting
005	Mitchell	HTTPS://OSF.IO/4XDKK/	Intermixed-blocked effect in perceptual learning
007	Beaman	HTTPS://OSF.IO/6N3BM/	Strategies & distributions of immediate memory
008	Dodson	HTTPS://OSF.IO/C5PBG/	Stereotypes & retrieval of illusory recollections
012	Marsh	HTTPS://osf.io/7rtcz/	Sequence phonological similarity, sound disruption
015	Schmidt	HTTPS://osf.io/bscfe/	Stroop, proportion congruence, and contingency
020	Sahakyan	HTTPS://OSF.IO/BZDR2/	Intentional forgetting after 1 or 2 "shots"
022	Colzato	HTTPS://OSF.IO/P9THW/	Bilingualism, executive control, inhibition
025	Couture	HTTPS://OSF.IO/K9GP6/	Corrects and errors in Hebb repetition effect
029	Turk-Browne	HTTPS://OSF.IO/UJHLW/	Multidimensional visual statistical learning
036	Pacton	HTTPS://OSF.IO/VMZ2E/	Attention-based account dependency learning
037	Makovski	HTTPS://OSF.IO/0PXRO/	Orienting attention, memory probe interference
106	Dessalegn	HTTPS://OSF.IO/IAJP5/	Language role in binding feature conjunctions
133	Nairne	HTTPS://OSF.IO/JHKPE/	Adaptive memory & value of survival processing
136	Vohs	HTTPS://OSF.IO/I29MH/	Determinism belief, cheating
158	Goschke	HTTPS://OSF.IO/BK53T/	Response conflict, prospective memory, cue monitor

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26978v1 | CC BY 4.0 Open Access | rec: 5 Jun 2018, publ: 5 Jun 2018

Table 2 (on next page)

Goodness of Fit for 18 data sets

Design, number of participants, maximum number of opportunities and BIC Goodness of Fit for probit and logit link analyses.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26978v1 | CC BY 4.0 Open Access | rec: 5 Jun 2018, publ: 5 Jun 2018

Table 2

Goodness of Fit for 18 data sets

ID	Design	N participant	N max	Probit BIC	Logit BIC	Logit /Probit
002	r4r2	24	24	202	306	1.51
003	r4	32	72	8	125	15.63
004	b3b2r2r2	30	24	946	1732	1.83
005	b2r2r2	24	8	335	518	1.55
007	Hr2	15	320	14	40	2.86
008	b2r2r2	24	32	512	686	1.34
012	b2r2r3	59	15	1202	2038	1.70
015	r3	242	144	-169	537	-3.18
020	b2r2	47	8	322	538	1.67
022	r2r2	32	30	139	262	1.88
025	r4	16	16	51	120	2.35
029	r2	30	16	57	84	1.47
036	b2r2r2	12	4	256	346	1.35
037	r2r2r2	24	4	166	371	2.23
106	b2	16	30	53	96	1.81
133	b2r2	19	16	123	191	1.55
136	b2	29	20	416	472	1.13
158	b2r2r2	7	18	803	1228	1.53

Notes. BIC = Bayesian Information Criterion, Logit/Probit = (logit BIC)/(probit BIC)

r is repeated, b is between factor, numbers after b/ or r are number of levels

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26978v1 | CC BY 4.0 Open Access | rec: 5 Jun 2018, publ: 5 Jun 2018