

Behavioral persistence in captive bears: a critique

Andrew R. Criswell^{1,3} and Gary J. Galbreath^{2,4}

¹Hedmark University College, Postboks 104,
N-2451, Rena, Norway

²Biological Sciences, Northwestern University,
Evanston, IL 60208, USA

Ursus 16(2):268–273 (2005)

Vickery and Mason (2003) sought to establish a causal relationship from stereotypy behavior to abnormal behavioral persistence in field experiments with 6 Asiatic black bears (*Ursus thibetanus*) and 6 Malayan sun bears (*Ursus malayanus*). For that study, stereotypy behavior was defined as "... actions that were repetitive, invariant, and without obvious function" (Vickery and Mason 2003:37). Abnormal behavioral persistence was characterized by bears who continued to respond to a previously learned, rewarded task after having been deprived of that reward. Among 12 bears studied, Vickery and Mason reported a statistically significant correspondence between abnormal behavioral persistence and stereotypy behavior. Furthermore, they claimed that among the 29 bears from their first-year study, time spent in captivity correlated with greater degrees of stereotypy behavior and thus greater aberrant behavior.

These findings led Vickery and Mason, in turn, to question the potential success of reintroduction programs with bears held captive for substantial time periods. "In the wild, where behavior must be adaptive and flexible to meet fluctuating conditions, such behavioral deficiencies could help account for reduced survivorship of reintroduced subjects" (Vickery and Mason 2003:35). In their view, the presence of stereotypy behavior and its associated abnormal behavioral persistence might render captive bears unable to adapt quickly to environmental change when reintroduced to the wild.

Reintroduction is such a potentially important conservation tool that findings questioning its efficacy or limiting its implementation must be carefully examined. With this in mind, we have scrutinized Vickery and Mason's results and report here our serious concerns regarding their methods and the conclusions derived from them. We first claim that the relationship

between stereotypy behavior and abnormal behavioral persistence Vickery and Mason found in their data is spurious. Second, we claim that their experimental hypothesis postulates a direction of causality between abnormal behavioral persistence and stereotypy behavior that is opposite to the statistical model they estimated and tested: they reversed the direction of causality. In a correlational study, that switch is not so serious, but in an experimental study such as theirs, it is crucial.

We are not taking issue with the enormous literature prior to Vickery and Mason's publication (Vickery and Mason 2003) which reports findings that stereotypy behavior develops in captive bears over time, that it has detrimental consequences, or that captive bears have fared poorly in reintroduction schemes. What concerns us here are the serious flaws in scientific method and statistical inference we found in Vickery and Mason (2003). We demonstrate here that the claims the authors make are unsupported by their data.

To clarify concepts and set notation, we denote the covariates of the model as E_i = number of trial responses taken to extinguish previously learned responses, L_i = number of trials to learn correct response, C_i = time in captivity (sometimes referred to in Vickery and Mason [2003] as "age"). T_i and G_i are each categorical variables denoting species and sex, respectively. The response variable is denoted as S_i = stereotypy frequency for bears $i = 1, \dots, n$ computed as the proportion of total blind-based observations for which stereotypy behavior was observed and measured "immediately prior to task presentation" (Vickery and Mason 2003:37). Another measure of stereotypy frequency, S'_i , is the mean stereotypy frequency measured over all observation periods.

These response data are binomial, continuous on a range between 0 and 1. As we argue, the authors failed consider this when they modeled the data with a linear specification. Table 1 reports the raw data from the Vickery and Mason (2003) study.

Learning extinction, captivity, and stereotypy behavior

Vickery and Mason (2003) estimated by least squares an equation of the form

$$\ln S_i = \beta_0 + \beta_1 E_i + \beta_2 L_i + \beta_3 C_i + \beta_4 T_i + \beta_5 G_i + \beta_6 (G_i \cdot T_i) + \varepsilon_i \quad (1)$$

which follows explicitly from Vickery and Mason (2003) (our Table 1, equation 2-1).

³andrew.criswell@osir.hihm.no ⁴gjj853@northwestern.edu

Table 1. Raw data used in Vickery and Mason's (2003) study of stereotypy in captive Asiatic black bears and Malayan sun bears (Provided to us by the authors).

Case ^a	T	G	C	E	L	S'	N ^b	SR ^c	S ^d
1	sun	female	8.0	41	140	0.45	248	99	0.40
2	sun	male	9.0	69	139	0.20	248	46	0.19
3	sun	female	7.0	60	120	0.18	247	50	0.20
4	sun	male	10.0	85	100	0.35	248	73	0.29
5	sun	male	9.0	48	140	0.15	248	35	0.14
6	sun	male	10.0	32	100	0.22	248	49	0.20
7	black	female	7.0	69	99	0.10	247	27	0.11
8	black	male	9.0	17	59	0.26	246	68	0.28
9	black	male	8.0	115	80	0.27	246	86	0.35
10	black	female	7.0	15	79	0.02	248	7	0.03
11	black	male	11.0	11	68	0.51	248	34	0.14
12	black	female	11.0	25	49	0.12	248	29	0.12
13	black	female	2.0	—	—	0.03	—	—	—
14	black	female	8.0	—	—	0.00	—	—	—
15	black	male	6.0	—	—	0.20	—	—	—
16	black	female	8.0	—	—	0.17	—	—	—
17	black	male	7.0	—	—	0.25	—	—	—
18	black	female	6.0	—	—	0.12	—	—	—
19	black	female	6.0	—	—	0.01	—	—	—
20	black	female	7.0	—	—	0.02	—	—	—
21	sun	female	2.0	—	—	0.12	—	—	—
22	sun	female	1.5	—	—	0.18	—	—	—
23	sun	male	8.0	—	—	0.37	—	—	—
24	black	male	4.0	—	—	0.02	—	—	—
25	black	male	8.0	—	—	0.02	—	—	—
26	sun	female	6.0	—	—	0.34	—	—	—
27	sun	female	7.0	—	—	0.34	—	—	—
28	black	female	6.0	—	—	0.00	—	—	—
29	black	female	5.0	—	—	0.24	—	—	—

^aCases 1–29 were used by Vickery and Mason (2003) for their penultimate study, their first equation and our equation (2); cases 1–12 for their ultimate study, their second equation, and our equations (1) and (3).

^bTotal number of observations made to detect the presence of stereotypy.

^cNumber of observations for which stereotypy was present.

^d $S = SR / N$ rounded to 2 decimal places, from Vickery and Mason (2003). Our estimates for stereotypy frequency rounded to 6 decimal places to minimize rounding errors. Results differ slightly.

We contend a model such as this, with 12 observations and complex interactions, is profoundly over-parameterized. The most extreme case of an over-fitted model is one in which the number of explanatory variables exhaust the data points. It does a perfect job of explaining that particular set of data but is useless in explaining anything beyond it. The point of statistical inference is one of generalization: to the degree a model is over-parameterized, generalization to the population at large is compromised (Nelson 1973).

Vickery and Mason claimed from equation (1) that “a significant positive relationship was found between an individual's stereotypy frequency and the number of responses made during extinction ...” (Vickery and Mason 2003:39). They failed, however, to report that their model as a whole fails to significantly explain variation in stereotypy frequency (Table 2, 2-1: $F = 3.128$; 6, 5 df; $P = 0.116$). We formed a likelihood ratio test to

determine whether exclusion of their main variable of interest, extinction, significantly reduces the model's fit. It did not (Table 2, equation 2-3 vs. 2-1: $F = 4.41$; 1, 5 df; $P = 0.090$). The authors would have obtained a far better fit had they chosen $\ln S'$ as the response variable in place of $\ln S$ (Table 2, equation 2-2: $F = 6.461$; 6, 5 df; $P = 0.029$). However, likelihood ratio tests for the exclusion of extinction again indicate it has little explanatory power, (Table 2, equations 2-4 vs. 2-3: $F = 3.22$; 1, 5 df; $P = 0.133$).

Time in captivity

Vickery and Mason (2003:39) argued that stereotypy frequency increased with age and gave expression to that view in terms of a regression equation,

$$S'_i = \beta'_0 + \beta'_1 C_i + \beta'_2 T_i + \beta'_3 G_i + \beta'_4 (G_i \cdot T_i) + \epsilon'_i, \quad (2)$$

Table 2. Regression estimates for persistence of stereotypy behavior in captive Asiatic black bears and Malayan sun bears for models used in Vickery and Mason (2003).

Dep. Var. Equation No.	Including extinction		Excluding extinction		S 2-5 (SE)	S 2-6 (SE)	E 2-7 (SE)
	ln S 2-1 (SE)	ln S' 2-2 (SE)	ln S 2-3 (SE)	ln S' 2-4 (SE)			
Intercept	-3.6348 (2.1046)	-6.5792 ^a (1.9247)	-2.0380 (2.4577)	-5.0337 ^a (2.1012)	0.0174 (0.0832)	0.2186 ^b (0.0449)	74.049 (144.67)
E	0.0110 (0.0053)	0.0086 (0.0048)					
C	0.1901 (0.1594)	0.4662 ^a (0.1458)	0.0693 (0.1862)	0.3718 (0.1592)	0.0266 ^a (0.0096)		-9.8909 (10.423)
L	-0.0014 (0.0107)	0.0102 (0.0098)	-0.0006 (0.0134)	0.0109 (0.0115)			0.1757 (0.7514)
G	-0.8574 (0.4185)	-1.1726 ^a (0.3827)	-1.1088 (0.5022)	-1.3692 ^a (0.4294)	-0.1183 ^a (0.0518)	-0.1431 (0.0574)	13.245 (36.441)
T	-0.2702 (0.6661)	-1.0985 (0.6091)	-0.1725 (0.8321)	-1.0221 (0.7114)	-0.0038 (0.0638)	0.0394 (0.0695)	14.397 (46.619)
T · G	1.6957 ^a (0.5751)	2.3301 ^b (0.5259)	1.6086 (0.7183)	2.2621 ^a (0.6141)	0.2336 ^a (0.0868)	0.1535 (0.0920)	-62.744 (53.519)
S							209.65 (135.3)
F	3.128	6.461	1.831	5.186	6.558	4.916	0.883
P-value	0.116	0.029	0.241	0.035	0.001	0.008	0.566
R ²	0.790	0.886	0.604	0.812	0.522	0.371	0.515

Standard errors in parentheses.

^aSignificant at 5% level.

^bSignificant at 1% level.

for which they found a significant relationship. "Time spent in captivity was found to be significantly associated with an individual's level of stereotypy, with older bears showing increased frequencies ($F = 7.59$; 1, 23 df; $R^2 = 52\%$; 1-tailed $P = 0.011$)." Note the distinction between the response variables Vickery and Mason (2003) employed in (2) and (1). We suggest it would have been better to remain consistent by using the same measure of stereotypy behavior. Furthermore, it was never made clear in their paper why 29 bears were employed to estimate (2) but only 12 bears for (1). Finally, it does not surprise us to see that Vickery and Mason did not apply a logarithmic transformation in this model although they did so for equation (1). Case numbers 14 and 28 (Table 1) have stereotypy frequencies of zero, for which a log transformation is undefined. This highlights the problem of applying linear models to a response variable which is fundamentally binomial, a point we take up next.

Misspecification of functional form

Many authors have chronicled the pitfalls in applying linear models to a binomial response variable (McCullagh and Nelder 1989, Christensen 1997, Agresti 2002, Collett 2003). For one, stereotypy frequency, although continuous, is bounded by 0 and 1, but estimation by

least squares pays little heed to that constraint, delivering predicted values that could easily exceed those bounds. For another, the assumption of constant variance is violated as the variance of residual error is specific to the various levels of stereotypy frequency in the sample. With equation (1) above, Vickery and Mason (2003) transformed stereotypy frequency by taking its natural logarithm, $\ln S_i$, "to meet the assumptions of parametric testing" (Vickery and Mason 2003:38). A logarithmic transformation hardly resolves the issue. An Anderson-Darling test failing to reject normality ($A^2 = 0.272$; $P = 0.602$) is but one diagnostic to base appropriate statistical treatment of the response variable. Examination of the predicted responses is yet another, and in this case, a more discerning criterion. As an example of the implausibility of their assumption, 5 out of 12 of the upper 95% prediction limits from (1) exceed unity for stereotypy frequency. For equation (2), even though the Anderson-Darling test fails to reject normality ($A^2 = 0.500$; $P = 0.192$), one of the predicted responses falls below zero (-0.05 for bear specimen 1), while 19 out of 29 of the lower 95% prediction limits are negative. To understand this phenomena better, we created 10,000 bootstrap resamples of the data which we used to estimate equation (1). We found that 23.5% of these bootstrap samples produced upper 95% prediction

Table 3. Logit regression estimates for data from Vickery and Mason (2003) on persistence of stereotypy behavior in captive Asiatic black bears and Malayan sun bears.

Dependent Variable	ln[S _i / (1 - S _i)]						
	Equation	3-1 ^a	3-2	3-3	3-4	3-5	3-6
Intercept		-2.7836 ^c (0.7981)	-2.7758 (2.7208)	-0.5432 (2.2412)	-1.4479 ^c (0.4057)	-1.0756 ^c (0.2875)	-1.4536 ^c (0.2793)
<i>E</i>		0.0103 ^c (0.0019)	0.0103 (0.0063)		0.0072 (0.0051)		
<i>L</i>		-0.0024 (0.0036)	-0.0024 (0.0123)	-0.0031 (0.0121)			
<i>C</i>		0.1445 ^b (0.0630)	0.1440 (0.2148)	-0.0340 (0.1800)			
<i>T</i>		-0.2849 ^c (0.2193)	-0.2853 (0.7484)	-0.1181 (0.7405)	-0.3381 (0.3972)	-0.2820 (0.3935)	0.3725 (0.3192)
<i>G</i>		-1.0353 (0.1663)	-1.0367 (0.5671)	-1.3161 ^b (0.5551)	-1.2050 ^b (0.5418)	-1.3030 ^b (0.5334)	-0.2834 (0.3285)
<i>T • G</i>		1.9627 ^c (0.2161)	1.9620 ^c (0.7373)	1.7944 ^b (0.7046)	1.7823 ^c (0.6874)	1.8172 ^c (0.6845)	
Scaled Deviance		59.338	5.101	7.830	5.915	7.901	15.325

Standard errors in parentheses.

^aDispersion parameter constrained to unity, otherwise free.

^bSignificant at 5% level.

^cSignificant at 1% level.

limits for S that exceeded unity. For equation (2), 53.2% of the bootstrap samples had predicted values for S' that fell short of zero.

For binomial response variables such as stereotypy frequency, "... calculated as a proportion of all observations (the number of observations of stereotypy divided by the total number of observations)" (Vickery and Mason 2003:37), the appropriate and commonly used procedure found in statistics textbooks on the subject (e.g., Fahrmeir and Tutz 2001, McCullagh and Nelder 1989) is one of logistic transformation, $\ln[S_i/(1 - S_i)]$, recommending use of generalized linear modeling technology.

We applied this model to the Vickery and Mason (2003) data (Table 3, equation 3-1) to calculate logit regression estimates of stereotypy frequency. The coefficient for extinction, 0.0103, while statistically significant ($z = 5.591$; $P < 0.001$), is biologically uninteresting. The estimated odds-ratio is $e^{0.101} = 1.010$ with 95% confidence interval ranging from 1.007 to 1.014. For a unit increase in the number of trials taken to extinguish behavior, the risk of stereotypy behavior rises an expected 1%.

Overdispersion

Equation (3-1, Table 3) does not account for extra binomial variation found in the response variable. The expected residual deviance for a well-behaved logit

model should equal its degrees of freedom (Collett 2003). The deviance for Equation (3-1) is 59.338 with 6 degrees of freedom, significantly different from its expected value ($D = 11.868$; 1 df; $P < 0.001$). Failing to account for overdispersion produces conservative standard errors and overstates the precision of estimated coefficients (Collett 2003).

Equations (3-2) through (3-6, Table 3) allow for inclusion of the dispersion coefficient estimated jointly for the baseline model (3-2) with no correction for overdispersion (Williams 1982, Collett 2003). It is clear from (3-2) that E , C and L are individually insignificant ($z = 1.637$, $P = 0.102$; $z = 0.670$, $P = 0.503$; and $z = -0.192$, $P = 0.848$, respectively). Furthermore, we fail to detect any significant deterioration in the fit of a model excluding E , C , L given interaction between sex and species, nor do we see any deterioration in a model that excludes only E given interaction between sex and species. The change in deviance between (3-2) and (3-5) forms a likelihood-ratio test for the exclusion of E , C , L given G , T . The loss in deviance through exclusion is clearly insignificant ($\Delta D = 2.800$; 3 df; $P = 0.424$). Comparison of (3-4) with (3-5) tests for the exclusion of E given G , T , whereas comparison of (3-2) with (3-3) allows for exclusion of E given E , L , G , T . In both cases, the loss in deviance is not significant ($\Delta D = 1.986$; 1 df; $P = 0.159$ and $\Delta D = 2.729$; 1 df; $P = 0.099$, respectively).

Thus, we conclude that when proper care is given to the correct functional form and allowance made for

overdispersion, the effect of initial learning, captivity, and learning extinction on stereotypy frequency disappears altogether.

Once E , C and L are recognized as explaining very little of the variation in stereotypy frequency, what remain are species and sex. Comparing the deviances of (3-5) with (3-6) reveals that interaction is important ($\Delta D = 7.242$; 1 df; $P = 0.006$). The fitted stereotypy frequency from (3-5) for the 3 female black bears is 0.085, significantly lower than for the 2 female sun bears, 0.301, for the 4 male sun bears, 0.205, and for 3 male black bears, 0.254 (Fig. 1). This is the only statistically significant relationship we find from Vickery and Mason's study: female Asiatic black bears displayed less stereotypy behavior than the rest.

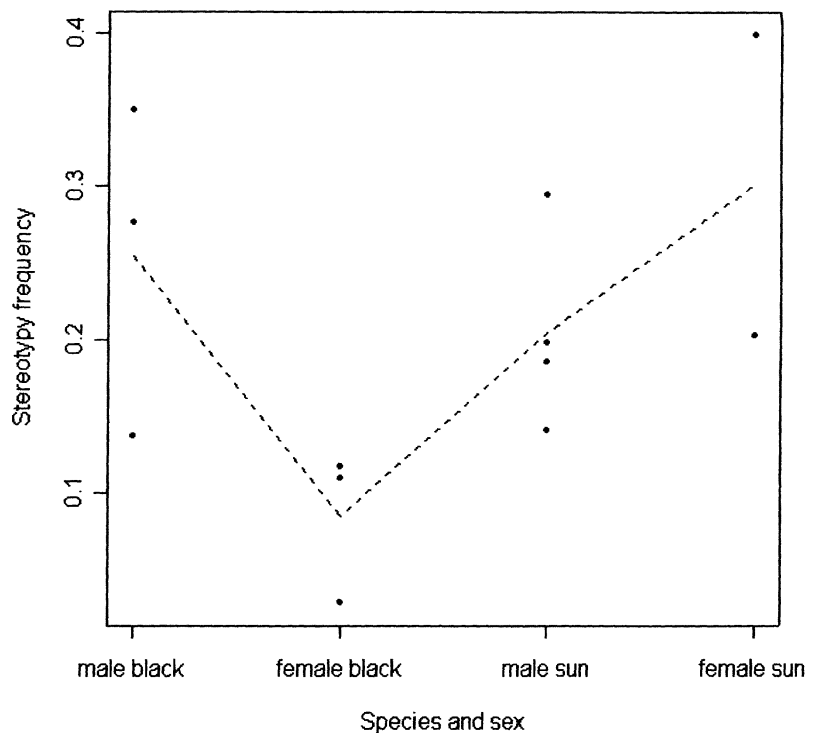


Fig. 1. Predicted stereotypy behavior of captive Asiatic black bears and Malayan sun bears. (Based on equation 3-5, Table 3).

Confusion over direction of causality

Vickery and Mason precisely stated their research hypothesis: "We hypothesized that bears displaying higher levels of stereotypy would take longer to extinguish responding ..." (Vickery and Mason 2003:38) and "... to test the hypothesis that, consistent with the finding in Garner and Mason (2002), animals with higher levels of stereotypy are behaviorally more persistent" (Vickery and Mason 2003:36). In other words, controlling for covariates C , L , G , T , an increase in stereotypy frequency should lead to an increased number of responses taken to extinguish previously learned behavior, $S \rightarrow E$. But the statistical model as specified in (1) is reversed: the direction of causality is $E \rightarrow S$. This error renders whatever empirical results Vickery and Mason provided as useless to the hypothesis they posed (see Appendix).

The proper specification, in accord with Vickery and Mason's hypothesis, is:

$$E_i = \beta_0'' + \beta_1''S_i + \beta_2''L_i + \beta_3''C_i + \beta_4''T_i + \beta_5''G_i + \beta_6''(G_i \cdot T_i) + \varepsilon_i'' \quad (3)$$

Equation (2-7, Table 2) reports estimates of (3). Overall, this equation fits the data poorly ($F = 0.883$; 6, 5 df; $P =$

0.566) with none of the covariates individually significant. In other words, nothing in Vickery and Mason's paper explains the extinction of learning behavior despite claims to the contrary. There is nothing in the data they collected from their experiments that supports the claim that "abnormal behavioral persistence in captivity could help explain why captive-bred bears often fare poorly after reintroduction" (Vickery and Mason 2003).

In a recent publication, Vickery and Mason (2005: 253) attempt to refute the criticisms lodged here with data on an additional 9 bears. However, these empirical findings are based on the very same flaws in scientific method and statistical inference employed in Vickery and Mason (2003) to which we have drawn the reader's attention.

Vickery and Mason's switching of $E \rightarrow S$ for $S \rightarrow E$ obfuscates the fundamental research question at issue: whether unmeasured neurological problems caused by captivity jointly affect stereotypy behavior and learning extinction. Taken either way, $E \rightarrow S$ or $S \rightarrow E$, we fail to detect, using their data, a statistically meaningful relationship between E and S . Our results show the fundamental research question they posed remains as it was before their publication.

Conclusion

Vickery and Mason (2003) lumped all the bears available to them into one treatment group, leaving for them nothing as a control group. This denied them the possibility of analyzing comparative treatment effects and the benefits of experimental inference. Their inference would have been stronger had they used an experimental design in which those bears that had previously been removed from enriched enclosures to small cages were compared to bears lucky enough to remain behind in the enclosures. Differential learning adaptation would indeed have been a study worthy of interest.

Studies of caged subjects cannot provide the entire picture about how captivity affects brain function and behavior because animals may adapt their behavior when placed in different environments. Enrichment programs may show that captive-held and captive-bred bears placed in a more favorable environment are likely to exhibit different learning adaptations than when housed in cages, a claim the authors support (Vickery and Mason 2003:41).

Reintroduction of captive-held or captive-bred animals to their natural habitat is a challenging task, with much of the literature to date attesting to those many varied difficulties. We question whether Vickery and Mason's paper constitutes a contribution to that debate. For the reasons given above, we conclude that Vickery and Mason's empirical findings are flawed, rendering their conclusions unsupported. The conservation community and government authorities charged with protecting wild animals and habitats rely on statements from the scientific community when making important decisions. For this reason, and for the sake of future wildlife conservation initiatives, we must ensure the statements made are accurate.

Acknowledgments

Special thanks go to P. Ratanaporn, G. van Zuylén, K. and J. Carter, T. Knutson, V. Ariyabudhiphongs, N. Ley, J. Fuller and M. E. Corey.

Literature cited

AGRESTI, A. 2002. *Categorical data analysis*. Second edition. John Wiley, New York, New York, USA.

- CHRISTENSEN, R. 1997. *Log-linear models and logistic regression*. Second edition. Springer-Verlag, New York, New York, USA.
- COLLETT, D. 2003. *Modelling binary data*. Second edition. Chapman and Hill, London, UK.
- FAHRMEIR, L., AND G. TUTZ. 2001. *Multivariate statistical modeling based on generalized linear models*. Springer-Verlag, New York, New York, USA.
- GARNER, J.P., AND G.J. MASON. 2002. Evidence for a relationship between cage stereotypes and behavioural disinhibition in laboratory rodents. *Behavioural Brain Research* 136:83–92.
- MCCULLAGH, P., AND J.A. NELDER. 1989. *Generalized linear models*. Chapman and Hill, London, UK.
- NELSON, C. 1973. *Applied time series analysis for managerial forecasting*. Holden Day, San Francisco, California, USA.
- VICKERY, S., AND G. MASON. 2003. Behavioral persistence in captive bears: implications for reintroduction. *Ursus* 14: 35–43.
- , AND ———. 2005. Stereotypy and perseverative responding in caged bears: further data and analyses. *Applied Animal Behaviour Science* 91:247–260.
- WILLIAMS, D. 1982. Extra binomial variation in logistic linear models. *Applied Statistics* 31:144–148.

Appendix

As a simple example to illustrate the problem involved with reversing causal direction, suppose the “truth” is $y = \beta x + \eta$ where y , x and η are each normally distributed with means zero and variances σ_y^2 , σ_x^2 and σ_η^2 , respectively. Suppose further that $E[x|\eta] = 0$. This establishes that $\beta = \sigma_{xy}/\sigma_x^2$. The fundamental research question at issue for a researcher unarmed with the truth but equipped with the proper specification and a sample of observations is one of independence, whether $E[y|x] = E[y]$ as inferred by that sample.

If the researcher erroneously believes $x = \alpha y + \varepsilon$ with $E[y|\varepsilon] = 0$, she or he would draw conclusions based on the estimated sample counterpart to $\alpha = \sigma_{xy}/\sigma_y^2$. The coefficient, α , differs from the “truth” by a ratio that depends on the relative variability of x to y , i.e., $\alpha = \beta\sigma_x^2/\sigma_y^2$. The bias in the magnitude but not direction of inference is a pure artifact of this relative variability. The significance of any inference, or lack thereof, especially in small samples as given here, is thus akin to the flip of a coin. A multivariate extension, more in line with Vickery and Mason's model, follows readily.