

Scaling the Critics

Uncovering the Latent Dimensions of Movie Criticism with An Item Response Approach*

Michael Peress[†]

Arthur Spirling[‡]

July 4, 2008

Abstract

We study the critical opinions of expert movie reviewers as an item response problem. We develop a framework that models an individual's decision to approve or disapprove of an item. Using this framework, we are able to recover the locations of movies and ideal points of critics in the same multi-dimensional space. We demonstrate that a three dimensional model captures much of the variation in critical opinions. The first dimension signifies movie 'quality' while the other two connote the nature and subject matter of the films. We then demonstrate that the dimensions uncovered from our 'threshold utility model' are statistically significant predictors of a movie's success, and are particularly useful in predicting the success of 'independent' films.

Key words: *threshold utility model* *film* *ideal points*

*Please do not cite or circulate without permission: comments are very welcome. Excellent research assistance from Chris Tice and Edward Laird is gratefully acknowledged.

[†]Department of Political Science, University of Rochester. mperess@mail.rochester.edu

[‡]Department of Political Science, University of Rochester. spln@mail.rochester.edu

1 Introduction

For the year 2006, the Motion Picture Association reported that international revenues generated by its composite companies totalled some \$42.6 billion (Hollinger, 2007). This sum is on a par with the gross domestic product of Kenya for the same period. Clearly then, the movie industry is an important economic force both in the United States (\$24.3 billion revenue for 2006) and elsewhere (\$18.3 billion). Fulfilling a consumer-advisory rôle within this massive sector, movie *critics* are ubiquitous: reviews and recommendations for films can be found in many journalistic outlets like newspapers, magazines and online websites. Major studios apparently accord substantial influence to such critics,¹ fêting them with press kits, advance screenings and other perks, and then using (selected, positive) reviewers' opinions directly in the marketing of their product.² Quite apart from their significance to large film-making firms and the news media devoted to the entertainment-industry, there is considerable *academic* interest in critics' choices and decision-making processes. First, within the marketing literature, assessing and quantifying the influence of critical reception on the commercial success of film media has been an ongoing concern (see, e.g., Ainslie, Drèze and Zufryden, 2005; Eliashberg and Shugan, 1997; Neelamegham and Chintagunta, 1999). Modeling the behavior of critics directly would thus paint a more complete picture of the interrelationship between film characteristics and market performance. Second, film criticism—particularly when practised by those versed in *film theory*—is an important element of cultural studies, a discipline that seeks to systematically understand cultural phenomena in terms of their social, political and psychological causes and consequences. Hence, there is motive to explore the ways in which audiences 'receive' the motion picture medium (see, e.g., Blumer, 1933; Kracauer, 1957; Mulvey, 1975; Riesman, Denny and Glazer, 1968). By analyzing new data on hundreds of critical reviews this paper seeks to contribute to both these scientific endeavors.

¹As do film historians: Smith (1998), for example, names the critics Gene Siskel and Roger Ebert in his top 100 ranking of the most influential people in movie history.

²Sony Pictures went so far as to create a fictional critic—named David Manning—whose enthusiastic (and entirely fabricated) 'quotes' appeared on several of the studio's movie adverts circa 2001. The company was subsequently sued for "intentional and systematic deception of consumers" and forced to pay over \$1 million in restitution to disgruntled movie-goers (Elsworthin, 2005).

As we will describe in more detail below, the data are examples of item responses (see, e.g., Hambleton, Swaminathan and Rogers, 1991; Lord, 1980, for an overview). Our central concern is using psychometric measurement techniques—especially those derived from item response theory (IRT)—to uncover the latent traits that characterize both a large number of critics and the movies that they review. The data differ from traditional applications since the subjects here choose whether to ‘approve’ or ‘disapprove’ of a *single* item. Hence, our theoretical framework of actor behavior leads us to employ a statistical approach that differs somewhat from the two- or three-parameter logistic models commonly seen in social science applications like educational testing (e.g. Rasch, 1981) or legislator ‘ideal point estimation’ (e.g. Poole and Rosenthal, 1997; Clinton, Jackman and Rivers, 2004). Our framework—the ‘utility threshold model’—applies to movie criticism, and more generally, to approval or ordinal rating data. Outside of movie criticism, our estimator applies to a number of other important problems. Legislators choose whether or not to cosponsor legislation. In marketing, a panel of consumers may be given a set of products to rate, and latent characteristics of these products could be deduced from these ratings. In admissions processes to universities, officers decide whether or not to allow a potential student entry based on their qualities.

In our application our model compensates for possible self-selection bias in the reviewing process by estimating and controlling for critic-specific ‘fixed effects’ and spatial preferences when locating the movies in space.³ Intriguingly, we find that the ‘expert’ critics in our data set—and the movies themselves— are almost fully described by two or three latent dimensions: they pertain to ‘quality’, followed by a division of space between ‘nerds’, ‘jocks’ and ‘art-house’. These latter labels refer to types of consumers who might enjoy predominantly science-fiction, action adventure and deep (potentially disturbing), emotional movies respectively. We demonstrate that such reviews are good predictors of financial success for movie makers, especially for independent films with relatively narrow audiences.

The paper proceeds as follows: in Section 2, we describe the movie critic data in more detail.

³Notice that a simple fixed effects model of critics would correct for the ‘stingy-ness’ of critics, but not the individual tastes of critics in terms of recovering the movie parameters.

We also review some relevant statistical literature. In Section 3 we introduce behavioral assumptions about reviewer behavior and its attendant notation. We then derive the estimation procedure. Section 5 describes our results in terms of scaling, while Section 6 demonstrates their importance for predicting movie success in the industry. Section 7 concludes and discusses further avenues for research.

2 Data and Background

Until relatively recently, data on critic responses to movies was both widely-scattered and in no standard form: different media recorded reviews in multiple ways, from long discursive articles with implicit judgements, to spoken television or radio reports to summary ‘star’-system recommendations. It was thus extremely costly to collate critical opinions. Moreover, the analyst was typically required to either use a few ‘key’ reviewers as indicative of a larger audience, or laboriously recode responses in order to make them comparable. The advent of the internet, however, has changed matters. The commercial outfit *Rotten Tomatoes*, a website situated at <http://www.rottentomatoes.com>, collates both multiple reviews for any given movie, and codes each review—in terms of how positive or negative it was towards the film—using a common rating system. In particular, *Rotten Tomatoes* considers each film review by each different critic (of which more than 100 may exist for recent movies) and then denotes the opinion as ‘fresh’ (i.e. the critic recommends the film) or ‘rotten’ (i.e. the critic does not recommend the film). This information is available to the public.

To see how this information might be used, first let $c = 1, \dots, C$ index the critics and $m = 1, \dots, M$ index the movies. The data to be modeled is then this $C \times M$ matrix of observed ratings (coded by *Rotten Tomatoes*) by the C critics on the M movies. Let \mathbf{Y} denote this matrix, with $y_{c,m}$ being the rating of critic c on movie m . We will code $y_{c,m} = 1$ if the critic does *not* recommend the movie (i.e. it is ‘rotten’) and $y_{c,m} = 2$ if the critic *does* recommend it (i.e. it is ‘fresh’). If no coding is present for a critic on a movie, we will code $y_{c,m} = 0$ as missing. If we include *all* critics and *all* movies then C and M are both quite large. Since not all critics review all movies (and some review very few), the matrix \mathbf{Y} contains a large number of missing values. This is not *per se* a problem for a frame-

work, if we rely on assumptions common in the item response and ideal point estimation literatures.

Our database uses a very expansive definition of what it is to be a film critic. Individuals who submit only a handful of film reviews to online mailing lists are considered critics. To focus on the population of interest—expert reviewers—we restrict \mathbf{Y} to all critics who are members of the *National Society of Film Critics*. This organization holds a prestigious place within the movie reviewing world and consists of approximately 60 respected individuals, all of whom are elected to their positions. A nice additional feature of this partition of \mathbf{Y} is that the critics involved typically write and turn in their reports for publication at approximately the same time: there is little danger, for example, that critics respond to *each other's opinions* rather than their viewing experience. We included all such critics who reviewed at least 20 films and all films that received at least 50 reviews on *Rotten Tomatoes*. The resulting dimensions are $C \times M$ where $C \approx 50$ and $M \approx 1000$.⁴

As should be clear, the matrix \mathbf{Y} (and its relevant partitions) contains rows of ‘individuals’ responding in a dichotomous way to ‘items’ in its columns. If we wish to understand the latent traits possessed by both critics and movies, IRT seems a reasonable way to proceed. Common social science applications of IRT include educational testing and, in this case, the Rasch (1981) model—essentially a one-parameter logistic (1-PL) model—is well known. Denoting the (reading) ability of a test taker c as θ_c and the difficulty of the item m as δ_m , the probability of success on any particular item is estimated as

$$\Pr(y_{c,m} = 1|\theta) = F(-(\theta - \delta_c)). \quad (1)$$

where $F(\cdot)$ is the inverse-logit. A version of Birbaum’s (see van der Linden and Hambleton, 1997, 13) two-parameter logistic model has received widespread use in political science (see, e.g., Clinton, Jackman and Rivers, 2004; Poole and Rosenthal, 1997), where the data are legislators responding—voting ‘yes’ or ‘no’—to parliamentary bills. Suppose that all legislators inhabit some ‘ideological’

⁴In practice, the minimum number of reviews a movie received among the NSFC critics was 16, while the median number of movies each NSFC critic reviewed was 336.

space of dimension d , in which they have some most preferred ‘ideal point.’ Moving policy outcomes away from this point means a utility loss for the legislator according to some pre-specified function. The probability that a legislator with an unobserved ideal point $\theta_c \in \mathbb{R}^d$ votes ‘yes’ to a particular proposition m is estimated as

$$\Pr(y_{c,m} = 1|\theta) = F(-a_m(\theta - b_m)) \quad (2)$$

where a_m is the degree to which the bill discriminates between legislators and b_m is the location of the ‘yes’ alternative in space relative to that of the status quo (the ‘no’ alternative). Such 1-PL or 2-PL models can be fit using maximum likelihood or Bayesian methods.

Neither of these popular models or their derivatives (see van der Linden and Hambleton, 1997, 19–22) is a natural choice for the current problem. Recall that critics either approve (‘recommend’) or disapprove (‘do not recommend’) a movie: this is somewhat different to, say, education testing (where latent ability is assessed) or roll call voting (where the implicit comparison is with the status quo). As we will show later, while traditional IRT approaches produce a reasonable reduced form model for approval data, they do not correspond to a useful structural model. The consequence is that, with a traditional approach, we cannot interpret the individual specific parameters as ideal points in a multidimensional space as would be common in the legislator ideal point literature.

3 Model and Estimation Procedure

Recall that $y_{c,m} = 2$ if the movie receives a positive review, $y_{c,m} = 1$ if the movie receives a negative review, and $y_{c,m} = 0$ if the movie is not reviewed by the critic. We assume the utility critic c gets from movie m is given by,

$$u_{c,m} = -(\alpha_c - \delta_m)'W(\alpha_c - \delta_m) + \epsilon_{c,m} \quad (3)$$

where α_c is a vector of critic ideal points, δ_m is a vector of latent movie attributes and W is a semi-positive definite weighting matrix that determines the relative importance of the dimensions. We let $\alpha, \delta \in \mathbb{R}^D$ where D denotes the dimensionality of both the critic *and* movie space. For example, there might be three dimensions (i.e. $D = 3$) in which all movies and critics can be situated: perhaps the first dimension corresponds to ‘action-ness’, the second to ‘romance-ness’ and the third to ‘drama-ness’. A romantic-comedy would have a ‘low’ score on the first dimension, but be ‘high’ on the other two. A critic who likes romantic-comedies over all other types of films would appear to have a similar scaling: otherwise put, $\Pr(y = 2)$ for a particular critic is maximized at $\alpha_c = \delta_m$.

A critic gives a positive review if $u_{c,m} \geq \bar{u}_c$. This allows each critic to have her own utility threshold. We assume that $\epsilon_{c,m}$ are independent identically distributed normal random variables with variance 1. Using this, we obtain a ‘fresh’ recommendation if $u_{c,m} \geq \bar{u}_c$ or if,

$$\epsilon_{c,m} \geq \bar{u}_c + (\alpha_c - \delta_m)'W(\alpha_c - \delta_m). \quad (4)$$

The probability that this is the case is given by,

$$\begin{aligned} \Pr(y_{c,m} = 1) &= \Phi(\bar{u}_c + (\alpha_c - \delta_m)'W(\alpha_c - \delta_m)) \\ \Pr(y_{c,m} = 2) &= 1 - \Phi(\bar{u}_c + (\alpha_c - \delta_m)'W(\alpha_c - \delta_m)) \end{aligned}$$

Note that the ‘zero’-dimensional model is of interest as well. In this model, there are no *spatial* locations of either critics or movies to be estimated: movies are treated as homogenous entities and the only source of heterogeneity is critical response to *quality*. That is, with respect to Equation (3), $u_{c,m} = \epsilon_{c,m}$, in which case, from Equation (4), a critic delivers a ‘fresh’ rating if $\epsilon_{c,m} \geq \bar{u}_c$. Thus, the only difference between critics is that some are stingier with their praise than others.

Combining all of this information, we can write the log-likelihood function as follows,

$$\begin{aligned} \mathcal{L}^{C,M}(\bar{u}, \alpha, \delta) = & \sum_{c=1}^C \sum_{m=1}^M [1\{y_{c,m} = 1\} \log \Phi(\bar{u}_c + (\alpha_c - \delta_m)'W(\alpha_c - \delta_m)) \\ & + 1\{y_{c,m} = 2\} \log \{1 - \Phi(\bar{u}_c + (\alpha_c - \delta_m)'W(\alpha_c - \delta_m))\}]. \end{aligned} \quad (5)$$

Estimating the parameters of the model can be accomplished by maximizing (5). This is straightforward in principle, but a number of complications arise. First, this model involves a very large number of parameters— $K = C(D + 1) + MD$. Hence, estimating the parameters of the model involves a potentially intractable computational problem. Second, despite our restriction to the NSFC critics there is still some sparseness in the data: some movies have few reviews while some critics opine on few films. There is thus potential (perfect-) separation in the data. For these reasons, we use a penalized-likelihood approach (in the sense of Firth, 1993).⁵ That objective function takes the following form:

$$\mathcal{L}^{C,M} + \sum_{c=1}^C \lambda_\alpha (\alpha_c' \alpha_c) + \sum_{m=1}^M \lambda_\delta (\delta_m' \delta_m) \quad (6)$$

where $\mathcal{L}^{C,M}$ is as given in Equation (5) and $\lambda_\alpha > 0$ and $\lambda_\delta > 0$ are penalty terms. Notice that the contribution of the penalty terms in the objective function approaches zero as the sample size increases: this is because the likelihood term from Equation (5) involves a double sum while each component of the penalty involves a single term.

As we will explain, the specification is close enough to standard IRT estimation routines that *mutandis mutatis* efficient algorithms developed for estimating those problems can be applied here. The differences are large enough that we cannot directly apply the same estimation routines to our problems.

⁵We follow the spirit rather than the letter of Firth's suggestions: we do not use a penalization based on Jeffrey's priors and we are not *per se* interested in asymptotic refinements.

3.1 Relationship to Applied IRT

The estimator we propose is closely related, but is it not isomorphic, to the estimators used for item response theory. The multidimensional item response model supposes that the probability of observing $y_{m,c} = 1$ (or a correct response, or a ‘yea’ vote) is given by

$$\Pr(y_{c,m} = 1) = F(a_m + b_m' \theta_c). \quad (7)$$

where $F(\cdot)$ is symmetric around zero and continuous.⁶

Applications of item response theory to legislative voting rely on the following derivation (Poole and Rosenthal, 1997). Let $x_m \in \mathbb{R}^D$ and $z_m \in \mathbb{R}^D$ denote the locations of two alternatives (e.g. a bill and a status quo) and let $\alpha_c \in \mathbb{R}^D$ denote the ideal point of an individual. Suppose that individuals have quadratic utility functions, subject to a random disturbance term. An individual’s utility from each of these options is characterized as

$$u_{c,m}^x = -(\theta_c - x_m)'(\theta_c - x_m) + \epsilon_{c,m}^x, \quad u_{c,m}^z = -(\theta_c - z_m)'(\theta_c - z_m) + \epsilon_{c,m}^z.$$

Here, an individual’s utility from an item is decreasing in the distance between that item and his ideal point and $\epsilon_{c,m}^x$ and $\epsilon_{c,m}^z$ are stochastic disturbance terms. Let $y_{m,c} = 1$ if the individual c chooses item x_m . Assuming that the individual chooses the item that yields him greater utility, we can determine that,

$$\Pr(y_{m,c} = 1) = \Pr(u_{c,m}^x \geq u_{c,m}^z) = \Pr(z_m' z_m - x_m' x_m + 2(x_m - z_m)' \theta_c + \epsilon_{c,m}^x - \epsilon_{c,m}^z \geq 0)$$

Reparameterizing the model using $\alpha_m = z_m' z_m - x_m' x_m$, $b_m = 2(x_m - z_m)$ and $\epsilon_{c,m} = \epsilon_{c,m}^x - \epsilon_{c,m}^z$, we obtain the item response model in its usual form.

This result implies we can interpret θ_c from an item response model as the ‘ideal point’ of the

⁶A number of equivalent parameterizations exist.

individual in a multidimensional space. The parameters a_m and b_m characterize the multidimensional cutting line, which divides the space into a region where the individual is more likely to prefer x_m and a region where the individual is more likely to prefer z_m . The structural model outlined above is appropriate when individuals face the choice between two alternatives (as is the case in legislative voting), but is not appropriate when individuals must choose between approval or disapproval of a single item (as is the case in our application).

As noted, we can show that our framework is not isomorphic to an item response model of any dimension. Our estimator can be written as

$$\Pr(y_{c,m} = 1) = F(\bar{u}_c + (\alpha_c - \delta_m)'W(\alpha_c - \delta_m)) = F(\bar{u}_c + \alpha_c'W\alpha_c - 2\alpha_c'W\delta_m + \delta_m'W\delta_m) \quad (8)$$

We can set up a relationship between the two models by letting $\theta_c = (W^{\frac{1}{2}}\alpha_{c,1}, \dots, W^{\frac{1}{2}}\alpha_{c,D}, \bar{u}_c) + \alpha_c'W\alpha_c$, $a_m = \delta_m'W\delta_m$, and $b_m = (-2W^{\frac{1}{2}}\delta_{m,1}, \dots, -2W^{\frac{1}{2}}\delta_{m,D}, 1)$. We now have that the D -dimensional threshold model is isomorphic to a $D + 1$ -dimensional item response model where the last component of b is restricted to be equal to 1.

This arrangement suggests that we cannot differentiate between the threshold and item response models on the bases of *model-fit* alone. Otherwise put, we can always find a $D + 1$ -dimensional item response model which summarizes the data at least as well as the D -dimensional utility-threshold model, and we can always find a $D + 1$ utility threshold model which summarizes the data at least as well as a D -dimensional item response model. Instead then, the advantage of the utility-threshold model is that it posits an appropriate *structural* model for the data, which allows us to correctly interpret the estimated parameters.

If the data were characterized by a D -dimensional utility-threshold model, we would be able to successfully fit a $D + 1$ -dimensional item response model. The difficulty would come in interpreting θ_c and (a_m, b_m) . Note that θ_c would contain the same information as $(\alpha_{c,1}, \dots, \alpha_{c,D}, \bar{u}_c)$, but the estimates would not reveal which components of θ_c characterize the ideal points and which compo-

nents characterize heterogeneity in the thresholds.⁷

A major advantage of our technique for approval data is that we can recover critic and movie locations in the same multidimensional space, something which would be impossible if we applied the traditional item response estimator to approval data.⁸

3.2 Identification

As is usual with such models we must impose some restrictions on the parameters in order to ensure identification. In the case of that standard multi-dimensional item response problem, it is well known that θ_c must be constrained for $D + 1$ individuals. A similar solution emerges here.

Unsurprisingly, the parameters of the utility threshold model are only identified up to location and scale. Specifically, consider the reparametrization,

$$\bar{\alpha}_c = A\alpha_c + b, \quad \tilde{u}_c = \bar{u}_c, \quad \tilde{\delta}_m = A\delta_m + b, \quad \tilde{W} = (A')WA^{-1} \quad (9)$$

where A has full rank. It is straightforward to show that

$$F(\tilde{u}_c + (\tilde{\alpha}_c - \tilde{\delta}_m)' \tilde{W}(\alpha_c - \tilde{\delta}_m)) = F(\bar{u}_c + (\alpha_c - \delta_m)' W(\alpha_c - \tilde{\delta}_m))$$

for all c, m . This indicates that we can apply a linear transformation to the critic ideal points without changing the value of the log-likelihood function, provided we can alter the other parameters in the model. To achieve point identification, we can normalize the first $D + 1$ ideal points (without loss of generality). Denote these constraints as $\tilde{\alpha}_c = \alpha_c = w_c$ for $c = 1, \dots, D + 1$ and

⁷This problem occurs because of the rotational invariance present in item response models and utility threshold models, meaning that \bar{u}_c need not appear as the last element of θ_c .

⁸We note that we can recover intelligence scores and item difficulties in the same one-dimensional space when applying the two parameter item response model to test data. We can recover legislator ideal points and cutting planes in the same multidimensional space when applying the multidimensional item response model to binary choice data. We cannot recover item *locations* in the same space in the binary choice application because we cannot separately identify the distance between the items and the variance of the disturbance term for that item.

define $\Delta_c = w_c - w_{D+1}$ for $c = 1, \dots, D$. The utility threshold model is identified provided that the vectors $(\Delta_1, \dots, \Delta_D)$ span \mathbb{R}^D . Without loss of generality, we can constrain $w_{D+1} = 0$ and $w_c = e_c$ for $c = 1, \dots, D$ where e_c is a unit vector.

Below we formalize the claim that the model is point identified. We effectively show that once we constrain the ideal points of $D+1$ critics, we cannot alter the parameter space leaving the value of the log-likelihood intact, with any transformation (linear or nonlinear).

Proposition 1 *Suppose that $\alpha_c = e_c$ where e_c is a unit vector for $c = 1, \dots, D$ and $\alpha_{D+1} = 0$ and W is a symmetric and positive definite matrix. Suppose that the vectors $\{\delta_m - \delta_{m'}\}_{m,m'}$ span \mathbb{R}^D . Suppose that $F(\cdot)$ is strictly increasing. We claim that there does not exist some parameter vector $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W})$ for which $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W}) \neq (\alpha, \bar{u}, \delta, W)$ with $\tilde{\alpha}_c = e_c$ for $c = 1, \dots, D$, \tilde{W} symmetric and positive definite and $\tilde{\alpha}_{D+1} = 0$ such that,*

$$F(\tilde{u}_c + (\tilde{\alpha}_c - \tilde{\delta}_m)' \tilde{W} (\tilde{\alpha}_c - \tilde{\delta}_m)) = F(\bar{u}_c + (\alpha_c - \delta_m)' W (\alpha_c - \delta_m)) \quad (10)$$

holds for all c, m .

Proof: see Appendix A.

Otherwise put, we claim that the estimated parameter vector uniquely gives rise to the data seen in practice: there exists no other vector that could possibly be responsible for the data.

4 Implementation

As we described above, the utility threshold model bears a strong resemblance to the item response models popular in the psychometric and political science literatures. The practical approaches used there fall into two broad categories. Fixed effects estimators treat both the item characteristics and individual characteristics as parameters to estimate (Bock and Lieberman, 1970; Lord, 1980; Poole and Rosenthal, 1997; Martin and Quinn, 2001). Random effects and conditional fixed effects estimators either integrate out or concentrate out the item (or individual) characteristics. The

random effects and conditional fixed effects estimators are useful when there are a small number of individuals or items. The fixed effects estimators have the advantage of producing additional information, which in our case includes both the individual (critic) and item (movie) specific parameters. Hence we take this approach.

Both maximum likelihood (Poole and Rosenthal, 1997) and Bayesian (Martin and Quinn, 2001; Clinton, Jackman and Rivers, 2004) versions of the fixed effects estimator have been applied in the political science literature. Programs for implementing these estimators are widely available but they cannot be *directly* applied here since, as noted, the information we wish to garner is not forthcoming from a standard item-response model. The Bayesian estimator is easier to implement efficiently, and modifying the existing code would not be very difficult. Experience indicates that the maximum likelihood estimator is more difficult to implement, yet it is computationally more efficient, particularly when the dimensionality is large. Because computational efficiency was a chief concern, we choose to implement the maximum likelihood estimator. The main difficulty of this approach is that we must optimize a function over a large number of parameters. For example, in a four dimensional model, there are more than 6,000 parameters to estimate. This optimization problem would usually be infeasible, but the special form of the objection function makes it tractable. In particular, the form of the objective function means that we can compute the objective function, the gradient, and the Hessian in operations.⁹ Our implementation relies on the *Zig-Zag* algorithm that has been applied to estimate nonlinear fixed effects models (Heckman, 1981) and ideal point models (Poole and Rosenthal, 1991, 1997).

A second concern with the fixed effects maximum likelihood estimator is obtaining estimates of uncertainty. This requires inverting the information matrix which has previously been thought to

⁹To save space, rewrite (5) as $\mathcal{L}^{C,M}(\bar{u}, \alpha, \delta) = \sum_{c=1}^C \sum_{m=1}^M \psi(y_{c,m}; \bar{u}_c, \alpha_c, \delta_m)$ and then note the following: derivatives with respect to any parameter involve only a single sum e.g. $\frac{\partial}{\partial \bar{u}_c} \mathcal{L}^{C,M}(\bar{u}, \alpha, \delta) = \sum_{m=1}^M \psi(y_{c,m}; \bar{u}_c, \alpha_c, \delta_m)$. Cross-derivatives involve only a single term, e.g. $\frac{\partial^2}{\partial \bar{u}_c} \mathcal{L}^{C,M}(\bar{u}, \alpha, \delta) = \frac{\partial^2}{\partial \bar{u}_c \partial \delta_m} \psi(y_{c,m}; \bar{u}_c, \alpha_c, \delta_m)$. Combined, these results indicate that $\frac{\partial}{\partial(\bar{u}, \alpha, \delta)} \mathcal{L}^{C,M}(\bar{u}, \alpha, \delta)$ and $\frac{\partial^2}{\partial(\bar{u}, \alpha, \delta)^2} \mathcal{L}^{C,M}(\bar{u}, \alpha, \delta)$ can be computed in $O(MC)$ operations which is significantly less than the $O(M^2C^2)$ and $O(M^3C^3)$ operations that would otherwise be required to compute the gradient and Hessian, respectively.

be computationally infeasible (Poole and Rosenthal, 1991; Lewis and Poole, 2004). Fortunately, efficient linear algebra routines allow us to invert as large as 10,000 in a manner of hours.

5 Results

We estimated a series of models, from zero through eight possible dimensions. Purely statistical goodness of fit measures for the fixed effects model are generally unavailable, due to the difficulty of working with asymptotic theory for data sets with large numbers of actors and large numbers of items. Indeed, standard statistical procedures will tend to recover a very large number of statistically significant dimensions which makes for a less than concise ‘summary’ of the data. We thus consider a number of alternative measures: the objective function value (the value of the log likelihood), the percent of observations correctly predicted, and the geometric mean probability (the average probability of a correct prediction). We also calculate the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for each alternative. Table 1 displays these measures for the various models.

In Figure 1 we report the same results graphically. The left y -axis gives the scale for the likelihood

	$D = 0$	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$	$D = 7$	$D = 8$
Log-likelihood	-10428.3	-8464.5	-7448.8	-6616.0	-5920.9	-5364.91	-4989.4	-4648.5	-4447.4
Geo Mean Prob	53.0%	66.2%	71.1%	75.2%	79.1%	82.4%	84.7%	86.6%	87.9%
Tot Percentage Correct	60.5%	77.2%	80.3%	84.6%	88.7%	92.7%	95.2%	97.3%	98.6%
Akaike IC	20944.60	20183.00	21317.60	22818.00	24593.80	26647.82	29062.80	31547.00	34310.8
Bayesian IC	21281.83	32652.99	45920.34	59553.49	73462.05	87648.82	102196.6	116813.5	131710.1

Table 1: Goodness-of-fit statistics for each model (dimensions 0 through 8).

while the right y -axis reports the geometric mean probability and the total percentage correctly predicted. Precisely which is ‘best’ model here is not obvious: the AIC suggests the $D = 1$ model is

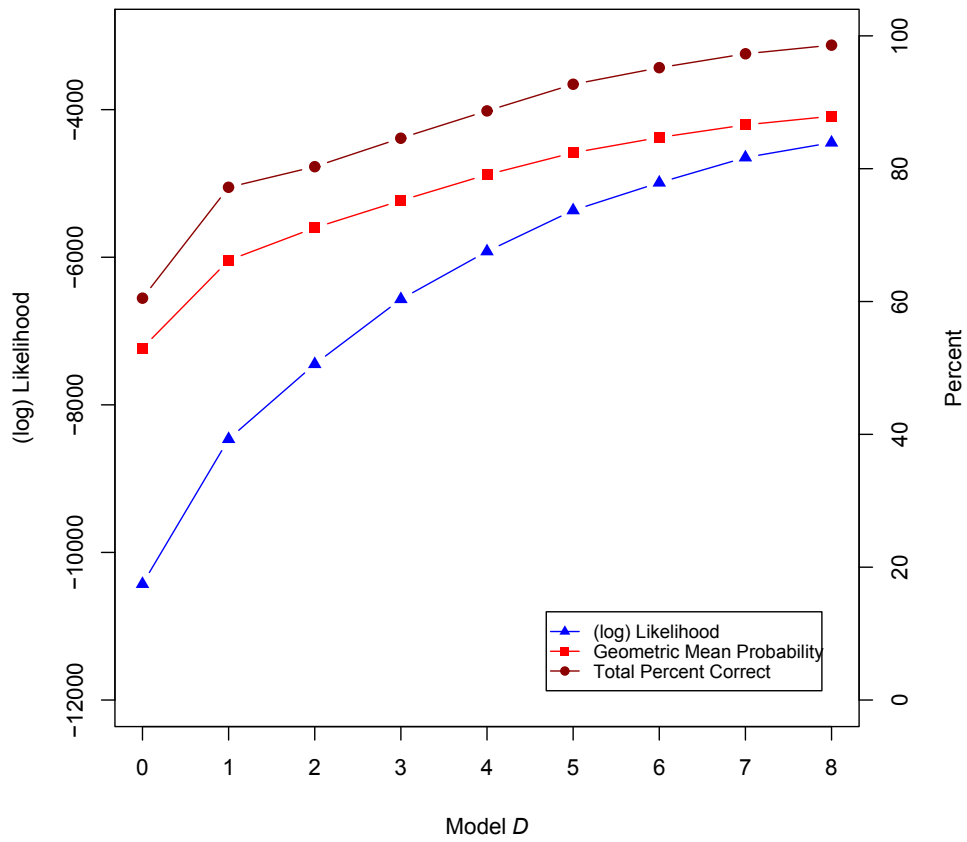


Figure 1: Goodness-of-fit for models of different dimensions. Left axis reports (log) likelihood, right axis reports geometric mean probability and total percent correct for each model.

preferred, while the BIC points to the zero-dimensional model. In terms of the percentage correctly predicted, the $D = 8$ model, unsurprisingly, does best. We thus choose a combination of complexity and fit: the $D = 3$ model predicts almost 85% of the cases correctly yet has an AIC comparable to the lower dimensional models. This is thus our preferred alternative and we interpret and discuss this version exclusively in what follows.

To recap, this model locates the movies and critics in three dimensions while also estimating the individual-level utility-thresholds for the critics. In Figure 2, we plot the density of the estimated utility-thresholds. Recall that a *lower* \bar{u} implies a more permissive critic who *ceteris paribus* is more willing to return a recommendation for the movie. From Figure 2, there is evidence of a slight negative skew: otherwise put, while the majority of critics are symmetrically located, there are a few ‘easily pleased’ individuals to the far left. Interestingly, the most generous critic is Roger Ebert (of the *Chicago Sun-Times*) who gives a ‘fresh’ rating 64% of the time. It is, by contrast, hard work to impress Amy Taubin, who writes columns for *The Village Voice*—she likes just 39% of the movies she reviews.

In Figure 3, we present a plot of the remaining three dimensions—and some movies and critics within this space. For the moment, we do not label the points, but they can be demarcated by their shape: the movies appear as round points, while the critics are triangles. A feature of Figure 3 is that the point clouds for critics and movies overlap, but not to the same extent in all dimensions.

To make this point clearer, consider Figure 4 where each of the respective two-dimensional plots is presented; again, the circular points represent the movies and the triangles are the estimated positions of the critics. In the top and middle panels, the movies and critics overlap much less than in the bottom panel. Otherwise put, the δ_1, α_1 dimension appears to discriminate between the groups in space. In particular, the critics generally appear to *right* of the movies: the critics have higher

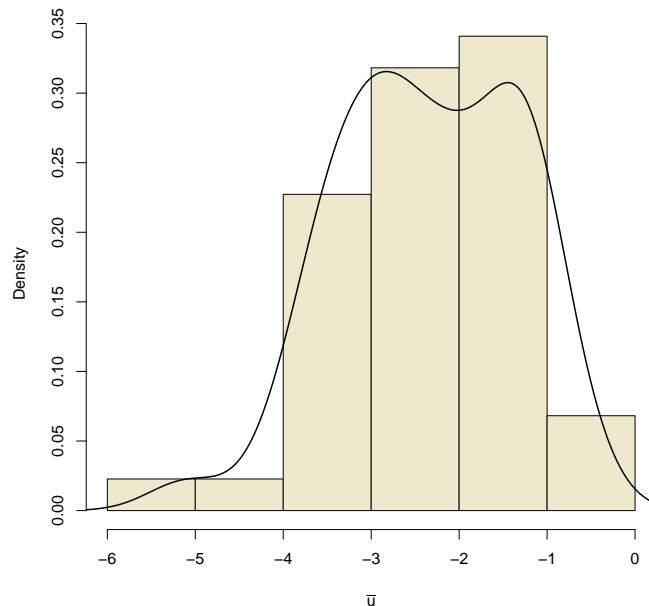


Figure 2: Histogram of utility thresholds for movie critics. Smoothed density superimposed.

estimated positions on this dimension.¹⁰ We contend that this dimension represents a movie’s ‘quality’ and, as we noted earlier, all else equal, critics prefer higher-quality movies to lower-quality ones. In our understanding, ‘high quality’ movies have a combination of two elements—artistic pre-tension and production values. Both refer to the craft and ingenuity of movie-making and we would expect ‘low quality’ movies to include so-called ‘B-movies’, pornographic and ‘exploitation’ films. As a more heuristic measure, we would contend that only *high* quality films would be in the running for an critics prize or Academy Award—be it for ‘Best Picture’, ‘Best Actor’, ‘Best Director’ and so on. With this in mind, large production budgets, impressive stunts and ‘A-list’ stars are certainly *not* a necessary condition for high quality, though they might generally be common features. In our conception, for ‘expert’ critics, quality is associated with the ‘high-mindedness’ of the movie as art, so small independent films could certainly be included within the rubric. High quality films might well be over-represented in certain genres such as romances, dramas and thrillers rather than, say,

¹⁰To be clear, under our original normalization, we discovered a dimension with a very high level of discrimination between critic and movie locations. We identified this as a quality dimension and rotated the data (exploiting rotational invariance) such that this dimension appeared as δ_1 , to aid in our interpretations.

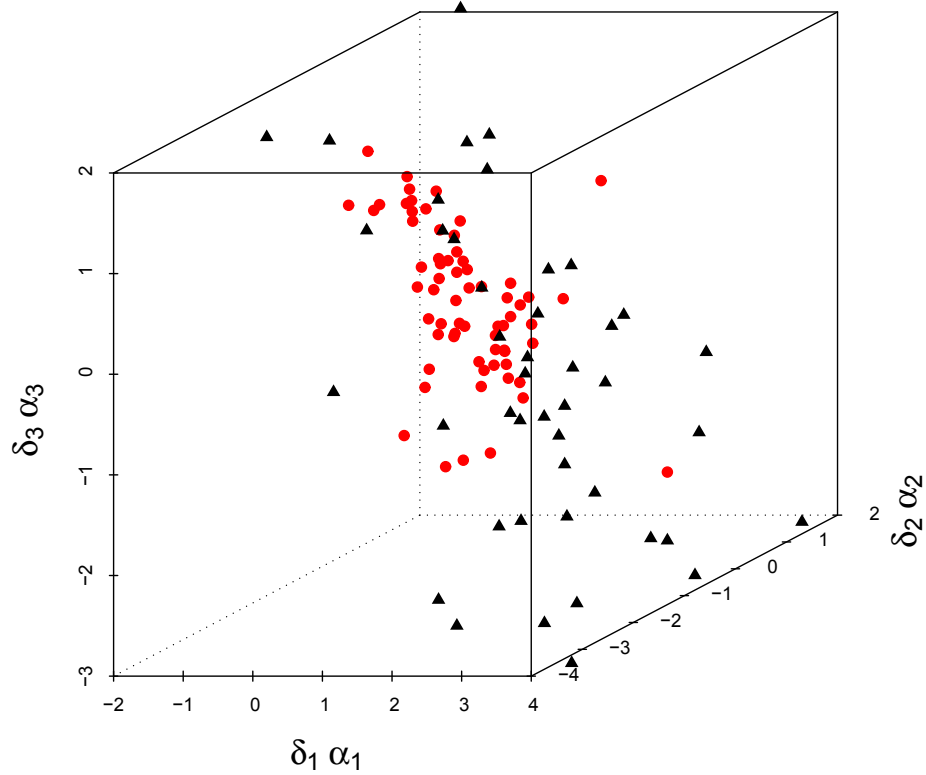


Figure 3: Scatter-plot for three dimensional spatial model. Circular points are movies; dark triangles are critics.

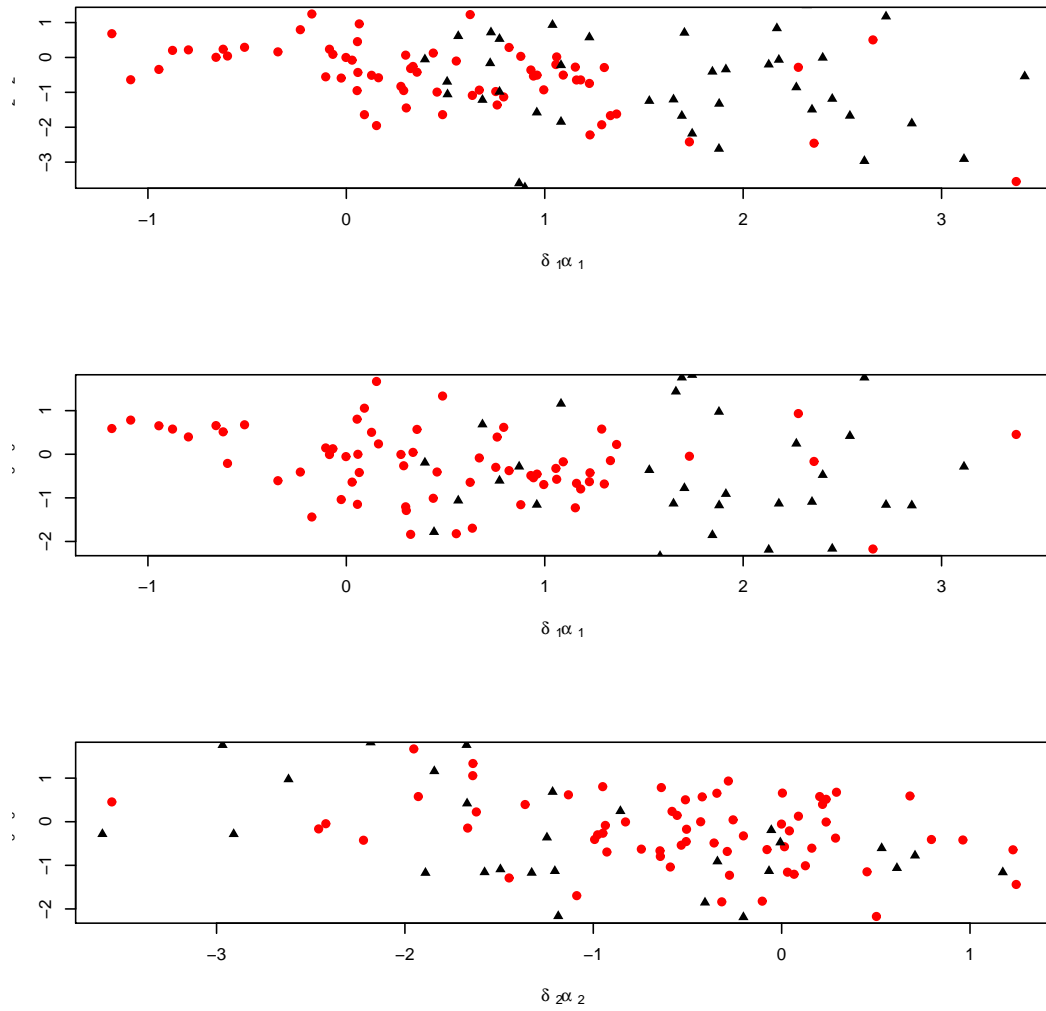


Figure 4: Scatter-plots for each of the three dimensions against the others. Movies are circular points, critics are dark triangles. Notice that the two groups show least overlap along the δ_1, α_1 access.

horror or action movies. We comment on this below. In Figure 5 we plot the density (and provide a histogram) of both the critics and movie estimates in δ_1, α_1 space—the dimension we claim is quality. Notice that there is some variance in the estimates for the critics; in our interpretation, this is due to sampling error rather than differing tastes for quality: *ceteris paribus* critics prefer high quality movies, but this does not mean that, say, a higher quality comedy is preferred to a lower quality drama.

Since we are sometimes dealing with relatively small numbers of reviews (e.g. *The Skeleton Key* of 2005 was reviewed by just four NSFC critics), there are reasonably large variances associated with our estimated movie qualities too. To avoid potentially misleading inferences then, in Table 2 we give some ranking information for the films in our sample at the 0.05, 0.5 (i.e. median) and 0.95 quantiles of their empirical cdf of the estimates for δ_1 . We also report the `rottentomatoes.com` aggregate (‘percent fresh’) rating for the movies and, in the final column, the genre description words given for the movies on the site. Notice that our δ_1 dimension estimates seem to agree with the aggregate ratings from the website; moreover, the genres seem fairly uniformly spread throughout the quality distribution, suggesting that this first dimension is indeed quality.

We now consider the other dimensions, δ_2 and δ_3 . Figure 6 is a scatterplot of selected movies in these two dimensions; for the purposes of the graphic we picked films for which we have a relatively large number of reviews (at least 18) and that covered the parameter space reasonably well but did not overlap with one another. This time, we label some of the movies. From Figure 6, we were not immediately able to identify the dimensions of film criticism. For example, *The Dreamers*, a French movie that deals with the sexual awakening of three teenagers during the strife of the 1968 Paris riots seems somewhat different in nature to *Alexander*, a huge budgeted historical epic starring Colin Farrell. Nonetheless these movies inhabit practically the same locations in space. We have similar concerns about the ‘closeness’ of *Melinda and Melinda*, which is a Woody Allen tragi-comedy set in Manhattan and *The War of the Worlds* which is a big-budget science-fiction fantasy film starring Tom Cruise.

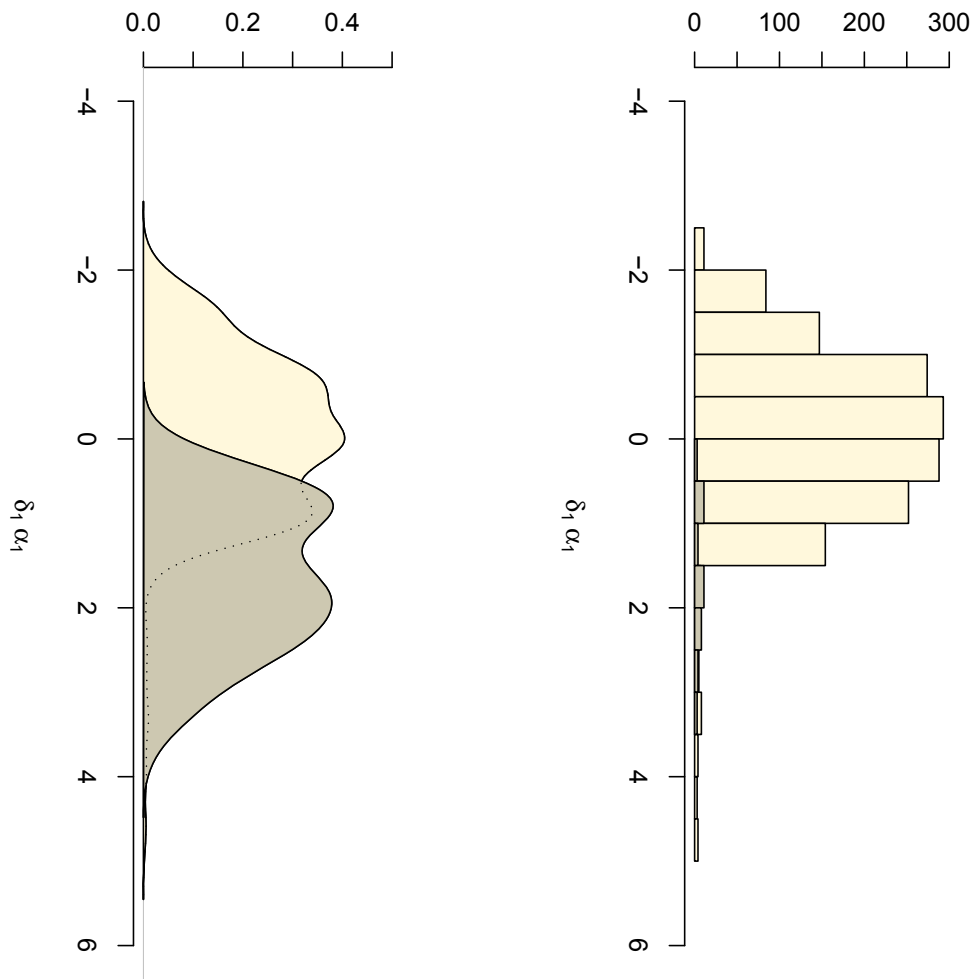


Figure 5: Histograms and (smoothed) density plots of movies (light color) and critics (dark color) in first dimension of model. We contend that this dimension is movie quality.

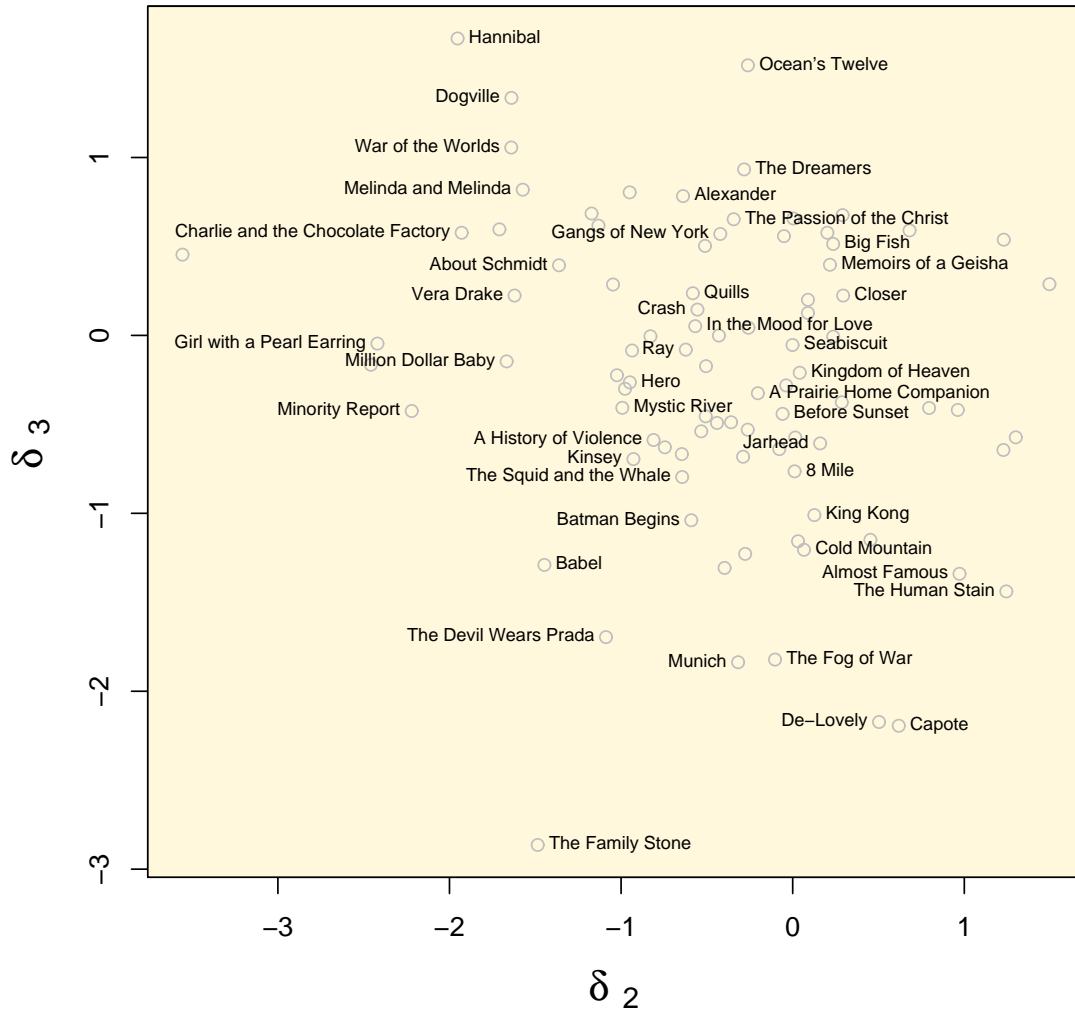


Figure 6: Scatterplot of movies in δ_2 and δ_3 space. Movies have 18 reviews or more.

Quantile	Title	Year	$\hat{\delta}_1$	% 'fresh'	Genre
0.95	Lost in Translation	2003	1.23	95	Dramas
	Kontroll	2005	1.223	81	Foreign Films
	Primer	2004	1.22	72	Dramas
	The Last King of Scotland	2006	1.22	88	Dramas
	This Film is Not Yet Rated	2006	1.208	84	Comedy
Median	Captain Corelli's Mandolin	2001	-0.08	28	Dramas
	Blood Work	2002	-0.08	56	Dramas
	Veronica Guerin	2003	-0.08	52	Dramas
	Hearts in Atlantis	2001	-0.08	48	Dramas
	The Low Down	2001	-0.07	60	Comedies
	Birth	2004	-0.07	39	Dramas
	Juwanna Mann	2002	-1.57	9	Comedies
0.05	Bulletproof Monk	2003	-1.58	22	Action/Adventure
	First Daughter	2004	-1.58	9	Comedies
	Jungle Book 2	2003	-1.58	20	Childrens
	Greenfingers	2001	-1.58	47	Dramas
	Dragonfly	2002	-1.58	7	Dramas

Table 2: Movies at and around the 0.05, median and 0.95 quantiles of the empirical CDF of $\hat{\delta}_1$. Final columns are *Rotten Tomatoes* aggregate rating and genre description from *Rotten Tomatoes*.

This finding initially seems troubling, but we suspect an explanation lies in the nature of the first, ‘quality’, dimension of movie review. Notice that Figure 6 does not distinguish (‘control for’) quality. Yet, put broadly, we would contend that ‘bad’ movies are actually very similar to one another: a bad comedy is not funny, a bad drama is not very dramatic, and a bad thriller does not leave one on the edge of the seat. Once these defining elements are removed, the movies appear almost identical, whatever one’s initial spatial preferences might have been. As an analogy, suppose one restaurant critic enjoys seafood, while another enjoys pasta-based meals. Also suppose that both are served multiple dishes of each type that are heavily over-salted. We suspect that the original (latent) preferences will be non-observable, because the critics will dislike everything they receive. Here then, we suspect that the failure to select on (high) quality works to disguise any spatial patterns in the data.

In Figure 7 we attempt to ameliorate this problem by presenting only those movies (with at least 15 reviews) that are ‘high’ quality. For present purposes this refers to those films that received

a δ_1 score above the 80th percentile of all values of δ_1 . In the figure, we also denote the (first) genre description of the movie as provided by *Rotten Tomatoes*, using different colors and plotting characters. We now note several patterns that were unapparent before. First, movies of a similar genre appear in groups, running broadly north-west to south-east across the plot. In particular, in the right, bottom corner, foreign films (open green triangles) cluster. North west of these come the dramas (filled red circles). Running in a north-south band to the west of the dramas are the comedies, interspersed with the action/adventure pictures. The science-fiction fantasy movies (black diamonds) appear to the west of the other movie types. In general, drama movies score relatively highly on δ_3 (and this is also true of foreign films), and have higher δ_2 values also. By contrast, science-fiction fantasy films are low on δ_2 while comedies are somewhere between the two. Comedies though, tend to have lower δ_3 scores. Action adventure movies are similar to comedies in this regard

To construct Figure 8, we took a different tack: here, the movies are colored and demarcated by their *Motion Picture Association of America* rating. As can be seen from the figure, the bulk of the ratings are either R, which denotes that any viewer under 17 years of age requires an accompanying parent or guardian, or PG-13 which denotes movies for which “Parents [are] Strongly Cautioned” and that might be inappropriate for children under 13 years of age. Broadly speaking, the R rated movies lie predominantly to the north and east of the PG and PG-13 movies which themselves run in a broad band from the west to the east and south of the graphic. As a result, the more family-friendly pictures tend to score lower on the δ_3 axis, and although they are somewhat similar regarding δ_2 . The ‘unrated’ movies help confirm this idea: generally lying to the north and east of the PG and PG-13 films, they include *Born into Brothels* which deals with the realities of child prostitution and *Capturing the Friedmans* which is a documentary concerning a father and son charged with child abuse. Presumably, neither of these films is suitable for minors.

Based on our assessment of Figure 7 and Figure 8, we present a combined graphic with our interpretation of the dimensions in Figure 9. We label the west of the graphic as ‘nerds’, denoting that movies in this area are popular among sci-fi fans. To the north-east of the plot, we denote

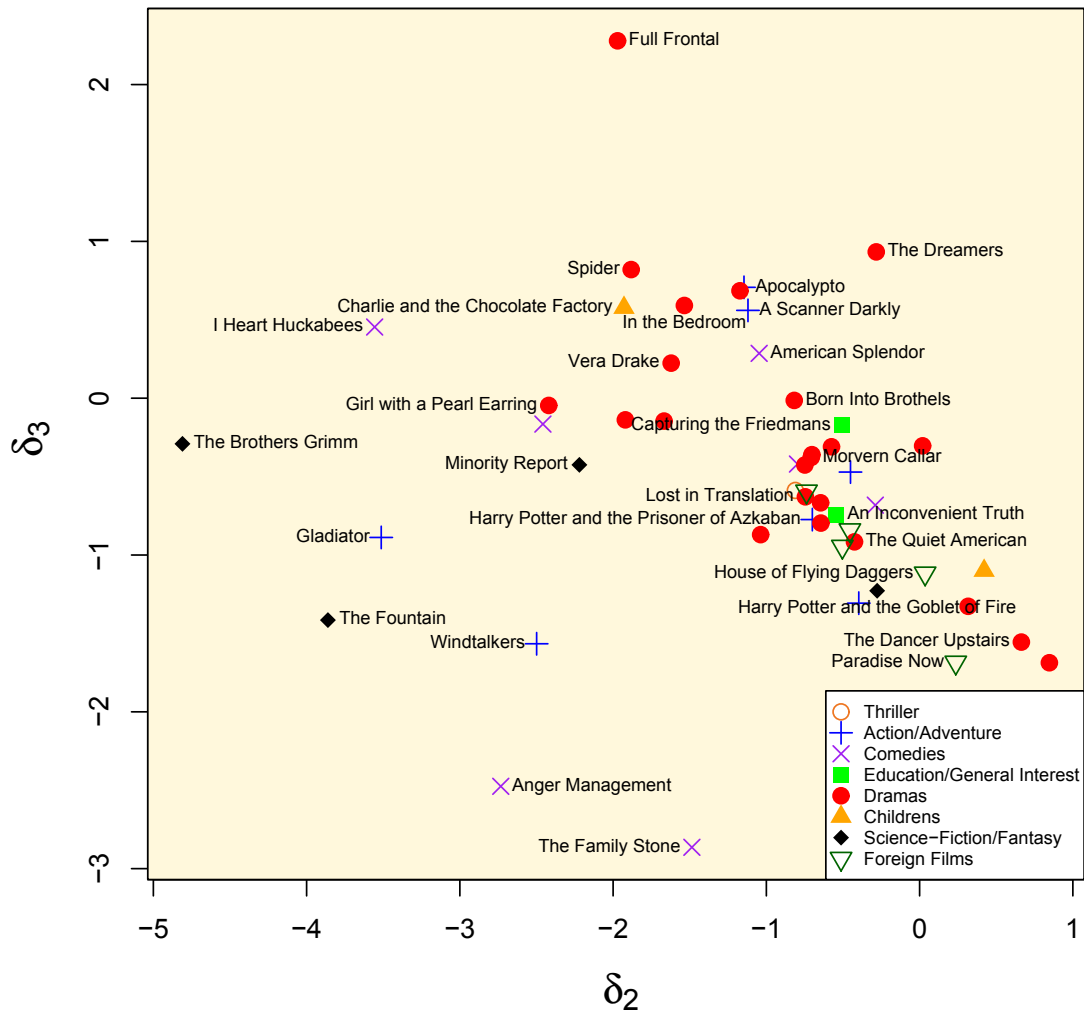


Figure 7: Scatterplot of movies in δ_2 and δ_3 space, plotting character and color denote genres. Movies have 15 reviews or more, and are 'high quality'.



Figure 8: Scatterplot of movies in δ_2 and δ_3 space, plotting character and color denote MPAA rating. Movies have 15 reviews or more, and are 'high quality'.

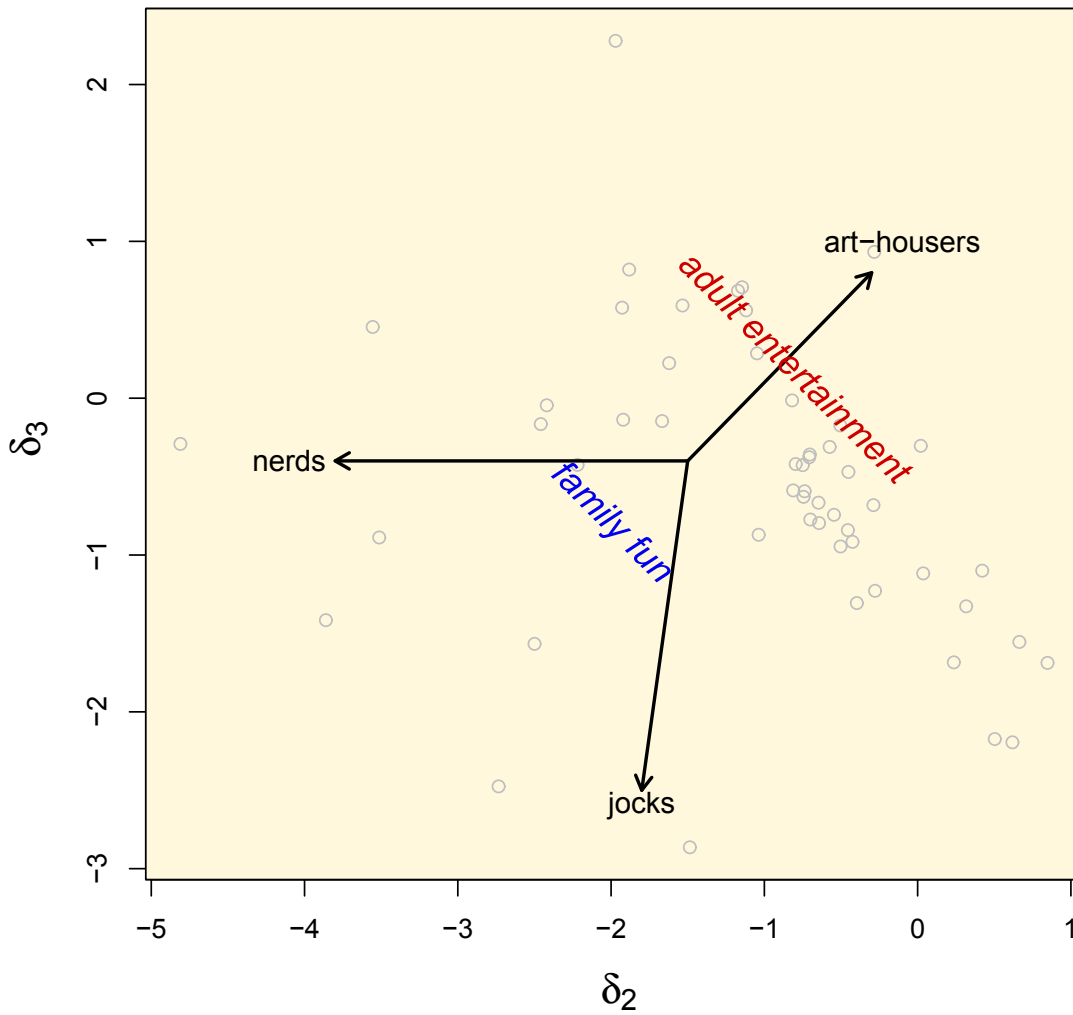


Figure 9: Scatterplot of movies in δ_2 and δ_3 space, with summary description. Movies have 15 reviews or more, and are ‘high quality’.

the area as ‘art-house’ to capture the fact that movies in this zone of the graphic might appeal to fans of (possibly pretentious, ‘deep’ and emotional) ‘art-house’ style pictures: *The Dreamers*, *In the Bedroom* and *Spider* all reside in this general direction. By contrast, to the south of the plot, we denote the area as ‘jocks’ and the movies here are predominantly action-adventure/comedy combinations: we think *Gladiator* and *Anger Management* would appeal to such fans. Overlaid on this plot are two descriptors that refer to the ratings of the movies: ‘adult entertainment’ refers (broadly) to films that receive at least an R rating, while ‘family fun’ refers to all other movies. Now that we have gone some way to establishing the dimensions of movie criticism, the next section analyzes the effects of these judgements on movie success.

6 The Effect of Movie Reviews

We believe that movie critics, via their reviews, have a perceptible effect on the success of movie performance. In this section we measure that performance as ‘profit’ which we define as the difference between (the log of) a film’s gross in the United States and the (log of) a film’s production budget.¹¹ Hence we assume that film-makers seek to maximize revenue minus costs. In addition to the reviews which are operationalized via our estimated $\hat{\delta}$, we have several other predictors to act as ‘controls’: `rating` which is a dummy for the MPAA rating the movie received; `create` which is a dummy denoting the creative type of the movie: ‘Contemporary Fiction’, ‘Factual’ and so on. We use a production type dummy (`prod.dum`) which includes categories like ‘live action’ or ‘stop motion animation’; a genre dummy (`genre.dum`) which denotes the movie’s primary genre, such as ‘drama’ or ‘romance’. We also record the movie’s *initial* release in terms of the number of screens it was shown at when opening (`init.theat`) and its ‘maximum’ release in terms of the total number of screens it showed on during its *entire* theater run (`max.theat`) as well as using a dummy (`holiday`) to account for possible profit variation due to the film’s opening falling on a holiday.

¹¹Data obtained from *The Numbers* website <http://www.the-numbers.com/> Mean of the dependent variable (profit) is approximately -0.07 , standard deviation is 1.35 .

	Estimate	Std. Error	t value	Pr(> t)
δ_1	0.1538	0.0573	2.68	0.0074**
δ_2	-0.0566	0.0549	-1.03	0.3031
δ_3	-0.0660	0.0516	-1.28	0.2008

Table 3: Coefficient table: OLS results for all movies. Dependent variable is profit (logged movie revenue minus logged movie costs). Details for controls available in Appendix B. Significance codes: ‘***’ $p < 0.001$, ‘**’ $p < 0.01$, ‘*’ $p < 0.05$, ‘.’ $p < 0.1$ Adj- R^2 is 0.32.

In Table 3 we report OLS results for our first model that includes all movies for which (complete) data is available; since the coefficients and other details on the controls are not of current interest, we drop them, though readers can view them in Appendix B. Interestingly, δ_1 is the only significant predictors for movie success. Recall that δ_1 is essentially movie quality, so a positive coefficient makes sense: the better the critics thought the movie was, the better it does at the box-office.

We were surprised to see that neither δ_2 (which we think is related to ‘nerdiness’) and δ_3 (which we think connotes ‘jockness’ and/or ‘art-houseness’) is significant. We suspected though, that NSFC critics are not to everyone’s tastes: they might not reflect the ‘general’ intended audiences for all the films. We thus split our sample into two parts: ‘wide-release’ movies that (by our definition) showed on *at least* 600 screens at the peak of their theater run, and ‘independent’ films that showed on *less than* 600 screens.¹²

In Table 4 we give the results for our wide-release regression: notice that now, δ_1 has a decreased p -value, and is no longer a predictor at the same significance level as before. This makes some sense if we regard the NSFC critics as being particular indicative of niche appeal. Table 5 confirms these ideas: we now see that all the components of the $\hat{\delta}$ estimate are significant at conventional levels for independent movies. Interestingly, ‘nerdiness’ (a low δ_2 value) is associated with more

¹²The industry standard defines a ‘wide-release’ as any film receiving an initial release of at least 600 screens. Problematically, some studios might release films for an initially ‘limited’ number of theaters to either (a) ensure their movie is eligible for Academy Awards (which requires it be released in a particular time frame for a given year) or to (b) ‘test the waters’ for a movie that might do poorly. We wanted to avoid counting such films as ‘independent’.

	Estimate	Std. Error	t value	Pr(> t)
δ_1	0.1790	0.0995	1.80	0.0730.
δ_2	0.0420	0.0917	0.46	0.6477
δ_3	0.0185	0.0861	0.22	0.8299

Table 4: Coefficient table: OLS results for ‘wide-release’ movies (maximum release ≥ 600 screens). Dependent variable is profit (logged movie revenue minus logged movie costs). Details for controls available in Appendix C. Significance codes: ‘****’ $p < 0.001$, ‘***’ $p < 0.01$, ‘**’ $p < 0.05$, ‘.’ $p < 0.1$ Adj- R^2 is 0.35.

	Estimate	Std. Error	t value	Pr(> t)
δ_1	0.1395	0.0709	1.97	0.0496*
δ_2	-0.1163	0.0693	-1.68	0.0936.
δ_3	-0.1248	0.0663	-1.88	0.0601.

Table 5: Coefficient table: OLS results for ‘independent’ movies (maximum release < 600 screens). Dependent variable is profit (logged movie revenue minus logged movie costs). Details for controls available in Appendix D. Significance codes: ‘****’ $p < 0.001$, ‘***’ $p < 0.01$, ‘**’ $p < 0.05$, ‘.’ $p < 0.1$ Adj- R^2 is 0.30.

profitable films, and in fact, the coefficient is larger than previously. Now too, δ_3 is a significant predictor, although we note that more ‘jock’ movies tend to do *better* at the box office (relative to ‘art-house’ movies).

Broadly speaking, our results imply that the NSFC critical reviews are either disproportionately influential in convincing independent movie fans, or disproportionately representative of them. Neither is particularly surprising: these critics are known for their expertise and presumably more ‘refined’ tastes (in the same sense that a restaurant critic will probably not recommend a fast food joint as his top choice), so we expect their views to resonate with more selective audiences.

Summary of Results

We have several findings from both Section 5 and Section 6, which may be summarized as follows:

- There are several dimensions of movie review for the NSFC critics: a three dimensional model provides a reasonable compromise between fit and complexity.

- Critics vary in terms of their ‘utility threshold’ for recommendation: whatever the nature of the movies, some are simply more ‘stingy’ than others.
- The first review dimension accords with movie ‘quality’: for all critics, ‘more’ is better.
- When low quality movies are included in the sample, substantive identification of review dimensions is very difficult
- When low quality movies are excluded, we notice a second ‘nerd’ dimension along with an ‘jock/art-house’ dimension to movie review.
- The NSFC opinions are statistically significant predictors of financial movie performance: movie quality is always positively associated with profit, but the other two dimensions are only important for *independent* films, released to a relatively limited audience.

7 Discussion

This paper developed a new ‘threshold utility model’ for estimating item response parameters of interest for movie critics and films they review. We argued that a three dimensional spatial model made most sense, and that the most important of these represented ‘quality’ of movies, for which, universally, ‘more’ is preferred to ‘less’. We presented evidence that such movie reviews are predictors of the financial success of movies, and that this effect is particularly strong for independent films.

In some IRT applications, notably educational testing, it makes sense to think of subjects and items in the *same* one-dimensional space: a test question has a particular ‘difficulty’ and a test-taker has an ‘ability’ on the same measurement line. In *multi*-dimensional, multi-parameter spatial models where the item receives a binary response—such as ‘ideal point estimation’ in legislatures—items and subjects cannot usually be placed in the same space. Such models typically have micro-foundations in which actors make pairwise comparisons between two available alternatives (say, the ‘status quo’ and a legislative proposal) and select their preferred option. This is clearly not

the case for critics: they choose to recommend a movie or not, without any attendant ‘default’ outcome. In light of this, we designed an approach with hybrid qualities: critics and movies *can* be located in similar (multi-dimensional) spaces and, to boot, we are able to estimate individual ‘quality’ thresholds for the critics.

There are several avenues for further research. Clearly, most consumer-advice critics operate in similar ways to our movie-reviewers: restaurants, books, paintings, exhibits and so on are ‘experienced’ and then a judgement passed. More broadly, most ‘satisfaction survey’-type exercises in marketing would yield data amenable to such analysis. We note that our framework can easily be extended to the case where individuals report multiple levels of satisfaction by incorporating more than one utility threshold. This would allow applications of our estimator to Likert scale data. In contrast to approaches relying on principal component analysis and related techniques, our estimator will produce estimates of product characteristics and rater ideal points in the same multidimensional space. In political science, promising applications include legislative cosponsorship and approval voting. Both of these have been studied to some degree using existing scaling techniques (Talbert and Potoski, 2002; Laslier, 2005), but we believe our approach can improve on these results by differentiating between spatial dimensions and heterogeneity in utility thresholds (following our argument in Section 3.1), and by providing estimates of the locations of bills and legislators, and voters and candidates, in the same multidimensional space.

A Identification of the Utility Threshold Model: Proof

Proposition 1 *Suppose that $\alpha_c = e_c$ where e_c is a unit vector for $c = 1, \dots, D$ and $\alpha_{D+1} = 0$ and W is a symmetric and positive definite matrix. Suppose that the vectors $\{\delta_m - \delta_{m'}\}_{m,m'}$ span \mathbb{R}^D . Suppose that $F(\cdot)$ is strictly increasing. We claim that there does not exist some parameter vector $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W})$ for which $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W}) \neq (\alpha, \bar{u}, \delta, W)$ with $\tilde{\alpha}_c = e_c$ for $c = 1, \dots, D$, \tilde{W} symmetric and positive definite and $\tilde{\alpha}_{D+1} = 0$ such that,*

$$F(\tilde{u}_c + (\tilde{\alpha}_c - \tilde{\delta}_m)' \tilde{W} (\tilde{\alpha}_c - \tilde{\delta}_m)) = F(\bar{u}_c + (\alpha_c - \delta_m)' W (\alpha_c - \delta_m)) \quad (11)$$

holds for all c, m .

Proof

We proceed by contradiction. We begin by supposing that there exists a $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W})$ with $\tilde{\alpha}_c = e_c$ for $c = 1, \dots, D$ and $\tilde{\alpha}_{D+1} = 0$ satisfying Equation (11). We will show that we must have $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W}) = (\alpha, \bar{u}, \delta, W)$. That is, that the condition requiring $(\tilde{\alpha}, \tilde{u}, \tilde{\delta}, \tilde{W}) \neq (\alpha, \bar{u}, \delta, W)$ is violated.

Since $F(\cdot)$ is strictly increasing, Equation (11) is equivalent to

$$\tilde{u}_c + (\tilde{\alpha}_c - \tilde{\delta}_m)' \tilde{W} (\tilde{\alpha}_c - \tilde{\delta}_m) = \bar{u}_c + (\alpha_c - \delta_m)' W (\alpha_c - \delta_m) \quad \forall c, m \quad (12)$$

Factoring out (12), we obtain

$$\tilde{u}_c + \tilde{\alpha}_c' \tilde{W} \tilde{\alpha}_c - 2\tilde{\alpha}_c' \tilde{W} \tilde{\delta}_m + \tilde{\delta}_m' \tilde{W} \tilde{\delta}_m = \bar{u}_c + \alpha_c' W \alpha_c + \delta_m' W \delta_m - 2\alpha_c' W \delta_m \quad (13)$$

$$\tilde{u}_c + \tilde{\alpha}_c' \tilde{W} \tilde{\alpha}_c - 2\tilde{\alpha}_c' \tilde{W} \tilde{\delta}_{m'} + \tilde{\delta}_{m'}' \tilde{W} \tilde{\delta}_{m'} = \bar{u}_c + \alpha_c' W \alpha_c + \delta_{m'}' W \delta_{m'} - 2\alpha_c' W \delta_{m'} \quad (14)$$

Subtracting (14) from (13) yields

$$-2\tilde{\alpha}'_c \tilde{W} \tilde{\delta}_m + 2\tilde{\alpha}'_c \tilde{W} \tilde{\delta}_{m'} + \tilde{\delta}'_m \tilde{W} \tilde{\delta}_m - \tilde{\delta}'_{m'} W \delta_{m'} = -2\alpha'_c W \delta_m + 2\alpha'_c W \delta_m - \delta'_{m'} W \delta_{m'} \quad (15)$$

When $c = D + 1$, we obtain

$$\tilde{\delta}'_m \tilde{W} \tilde{\delta}_m - \tilde{\delta}'_{m'} \tilde{W} \tilde{\delta}_{m'} = \delta'_m W \delta_m - \delta'_{m'} W \delta_{m'} \quad (16)$$

Plugging (16) into (15), we obtain

$$-2\tilde{\alpha}'_c \tilde{W} \tilde{\delta}_m + 2\tilde{\alpha}'_c \tilde{W} \tilde{\delta}_{m'} = -2\alpha'_c W \delta_m + 2\alpha'_c W \delta_{m'}. \quad (17)$$

When $c = 1, \dots, D$, Equation (17) yields

$$e'_c \tilde{W} (\tilde{\delta}_{m'} - \tilde{\delta}_m) = e'_c W (\delta_{m'} - \delta_m) \quad (18)$$

This further implies that,

$$\tilde{W} (\tilde{\delta}_{m'} - \tilde{\delta}_m) = W (\delta_{m'} - \delta_m). \quad (19)$$

Plugging Equation (19) into (17), we have

$$\tilde{\alpha}'_c W (\delta_{m'} - \delta_m) = \alpha'_c W (\delta_{m'} - \delta_m). \quad (20)$$

Since this must hold for all m , W has full rank, and the vectors $\{\delta_{m'} - \delta_m\}_{m, m'}$ span \mathbb{R}^D we have that

$$\tilde{\alpha}_c = \alpha_c \quad (21)$$

Now, we can use Equation (12) to determine that

$$\tilde{u}_c - \bar{u}_c + \tilde{\delta}'_m \tilde{W} \tilde{\delta}_m - \delta'_m W \delta_m = 2\alpha'_c \tilde{W} \tilde{\delta}_m - 2\alpha'_c W \delta_m \quad (22)$$

Using $c = 1, \dots, D + 1$ we obtain,

$$\tilde{u}_c - \bar{u}_c + \tilde{\delta}'_m \tilde{W} \tilde{\delta}_m - \delta'_m W \delta_m = 2e'_c \tilde{W} \tilde{\delta}_m - 2e'_c W \delta_m \text{ for } c = 1, \dots, D \quad (23)$$

$$\tilde{u}_{D+1} - \bar{u}_{D+1} + \tilde{\delta}'_m \tilde{W} \tilde{\delta}_m - \delta'_m W \delta_m = 0 \quad (24)$$

Subtracting (14) from (13), we obtain

$$\frac{1}{2}(\tilde{u}_c - \bar{u}_c - \tilde{u}_{D+1} + \bar{u}_{D+1}) + e'_c W \delta_m = e'_c \tilde{W} \tilde{\delta}_m. \quad (25)$$

Stacking these, and using the fact that \tilde{W} is nonsingular, we obtain,

$$\tilde{\delta}_m = \omega + \tilde{W}^{-1} W \delta_m \quad (26)$$

where

$$\omega = \frac{1}{2} \tilde{W}^{-1} \begin{pmatrix} \tilde{u}_1 - \bar{u}_1 - \tilde{u}_{D+1} + \bar{u}_{D+1} \\ \tilde{u}_2 - \bar{u}_2 - \tilde{u}_{D+1} + \bar{u}_{D+1} \\ \vdots \\ \tilde{u}_D - \bar{u}_D - \tilde{u}_{D+1} + \bar{u}_{D+1} \end{pmatrix}$$

Recall that Equation (16) implies that

$$2\omega' W (\delta_m - \delta_{m'}) = (\delta_m + \delta_{m'})' (W - W \tilde{W}^{-1} W) (\delta_m - \delta'_{m'}) \quad (27)$$

If $\delta_m \neq \delta_{m'}$, the result implies that either $W - W \tilde{W}^{-1} W$ or $\delta_m + \delta_{m'}$ is a constant for all m, m' .

We know that the second condition is not true, in which case $\tilde{W} = W$. This further implies that

$$\tilde{\delta}_m = \omega + \delta_m \quad (28)$$

Using Equation (12) once again, we have

$$\tilde{u} + \omega' W \omega - 2\alpha'_c W \omega + \omega' W \delta_m = \bar{u}_c. \quad (29)$$

This result implies that $\omega = 0$ and hence $\tilde{\delta}_m = \delta_m$. Using Equation (29), we have $\tilde{u} = \bar{u}_c$ and the result obtains. ■

B Full Table: OLS Results, all movies

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.6946	0.4662	-7.93	0.0000
Delta1	0.1538	0.0573	2.68	0.0074
Delta2	-0.0566	0.0549	-1.03	0.3031
Delta3	-0.0660	0.0516	-1.28	0.2008
rating1	0.2038	0.6117	0.33	0.7391
rating2	-4.7614	1.2954	-3.68	0.0003
rating3	-0.8290	0.5774	-1.44	0.1515
rating4	-0.0744	0.5205	-0.14	0.8863
rating5	-0.2956	0.5034	-0.59	0.5572
rating6	-0.3791	0.5057	-0.75	0.4537
createContemporary Fiction	1.9833	0.2151	9.22	0.0000
createDramatization	1.6125	0.2562	6.29	0.0000
createFactual	2.8786	0.5330	5.40	0.0000
createFantasy	1.4246	0.2758	5.17	0.0000
createHistorical Fiction	1.4963	0.2385	6.27	0.0000
createKids Fiction	1.7628	0.3065	5.75	0.0000
createScience Fiction	1.4044	0.2803	5.01	0.0000
createSuper Hero	1.9696	0.3585	5.49	0.0000
prod.dum1	1.7556	0.4317	4.07	0.0001
prod.dum2	1.8109	0.4681	3.87	0.0001
prod.dum3	1.0280	0.4701	2.19	0.0290
prod.dum4	1.3595	0.2949	4.61	0.0000
prod.dum6	0.7482	1.2188	0.61	0.5395
prod.dum7	1.8668	0.9072	2.06	0.0399
genre.dumAction/Adventure	0.5572	0.3501	1.59	0.1118
genre.dumAnimated	0.2242	0.6210	0.36	0.7181
genre.dumChildrens	0.5351	0.4172	1.28	0.2000
genre.dumComedies	0.8985	0.3384	2.66	0.0081
genre.dumDramas	0.6562	0.3391	1.94	0.0533
genre.dumEducation/General Interest	2.9881	0.6372	4.69	0.0000
genre.dumForeign Films	0.4118	0.5532	0.74	0.4569
genre.dumHorror/Suspense	1.3695	0.3613	3.79	0.0002
genre.dumRomance	1.1231	0.5009	2.24	0.0252
genre.dumScience-Fiction/Fantasy	0.6390	0.4364	1.46	0.1435
genre.dumThriller	0.6746	0.4663	1.45	0.1484
init.theat	0.0004	0.0002	1.69	0.0912
max.theat	-0.0001	0.0002	-0.30	0.7653
holiday1	0.2259	0.2236	1.01	0.3127
holiday2	0.1204	0.1847	0.65	0.5145
holiday3	-0.0318	0.1920	-0.17	0.8687
holiday4	0.3416	0.2456	1.39	0.1647
holiday5	-0.1984	0.3220	-0.62	0.5379
holiday6	-0.4815	0.2292	-2.10	0.0360
holiday7	0.1030	0.1828	0.56	0.5732
holiday8	0.0162	0.1780	0.09	0.9274
holiday9	0.1158	0.3214	0.36	0.7187
holiday10	0.0791	0.1866	0.42	0.6719

Table 6: Full coefficient table: OLS results for all movies. Dependent variable is profit (logged movie revenue minus logged movie costs).

C Full Table: OLS Results, ‘wide-release’ movies

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.1464	0.9816	-2.19	0.0297
Delta1	0.1790	0.0995	1.80	0.0730
Delta2	0.0420	0.0917	0.46	0.6477
Delta3	0.0185	0.0861	0.22	0.8299
rating1	-1.0089	1.8476	-0.55	0.5855
rating3	-2.4277	1.8058	-1.34	0.1800
rating4	-1.3929	1.7340	-0.80	0.4226
rating5	-1.5205	1.7252	-0.88	0.3789
rating6	-1.4477	1.7472	-0.83	0.4081
createContemporary Fiction	1.6809	0.3108	5.41	0.0000
createDramatization	1.2351	0.3872	3.19	0.0016
createFactual	3.6283	1.1585	3.13	0.0019
createFantasy	1.5092	0.4250	3.55	0.0005
createHistorical Fiction	0.8401	0.3652	2.30	0.0222
createKids Fiction	1.7551	0.4416	3.97	0.0001
createScience Fiction	1.4592	0.4576	3.19	0.0016
createSuper Hero	1.7676	0.4911	3.60	0.0004
prod.dum1	2.1982	0.6634	3.31	0.0011
prod.dum2	2.2520	0.6782	3.32	0.0010
prod.dum3	0.8103	0.7170	1.13	0.2595
prod.dum4	1.7652	0.5159	3.42	0.0007
prod.dum7	2.2501	1.4717	1.53	0.1275
genre.dumAction/Adventure	2.5011	1.6972	1.47	0.1418
genre.dumAnimated	2.2587	1.8150	1.24	0.2145
genre.dumChildrens	2.3956	1.7118	1.40	0.1629
genre.dumComedies	2.7076	1.6748	1.62	0.1072
genre.dumDramas	2.7461	1.6928	1.62	0.1060
genre.dumEducation/General Interest	3.0816	1.3406	2.30	0.0223
genre.dumForeign Films	1.4922	2.1547	0.69	0.4892
genre.dumHorror/Suspense	3.1309	1.7185	1.82	0.0696
genre.dumRomance	3.0342	1.7991	1.69	0.0929
genre.dumScience-Fiction/Fantasy	1.9432	1.7353	1.12	0.2639
genre.dumThriller	2.7581	1.8580	1.48	0.1389
init.theat	-0.0003	0.0004	-0.72	0.4733
max.theat	-0.0028	0.0008	-3.44	0.0007
holiday1	-0.0878	0.5314	-0.17	0.8690
holiday2	0.4048	0.5121	0.79	0.4300
holiday3	0.1657	0.4497	0.37	0.7128
holiday4	0.4267	0.4215	1.01	0.3123
holiday5	0.5842	0.5431	1.08	0.2831
holiday6	-0.8597	0.4890	-1.76	0.0799
holiday7	0.4697	0.4882	0.96	0.3369
holiday8	0.3019	0.3263	0.93	0.3558
holiday9	0.2232	0.5204	0.43	0.6684
holiday10	0.2090	0.3376	0.62	0.5364

Table 7: Full coefficient table: OLS results for wide-release movies (≥ 600 screens at maximum release). Dependent variable is profit (logged movie revenue minus logged movie costs).

D Full Table: OLS Results, ‘independent’ movies

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.0490	0.6393	-4.77	0.0000
Delta1	0.1395	0.0709	1.97	0.0496
Delta2	-0.1163	0.0693	-1.68	0.0936
Delta3	-0.1248	0.0663	-1.88	0.0601
rating1	-0.3142	0.9007	-0.35	0.7274
rating2	-5.7999	1.3820	-4.20	0.0000
rating3	-1.0973	0.7528	-1.46	0.1455
rating4	-0.6010	0.6769	-0.89	0.3750
rating5	-0.9106	0.6521	-1.40	0.1632
rating6	-0.9916	0.6527	-1.52	0.1293
createContemporary Fiction	2.1095	0.3178	6.64	0.0000
createDramatization	1.6963	0.3645	4.65	0.0000
createFactual	2.9475	0.6614	4.46	0.0000
createFantasy	1.3660	0.3853	3.55	0.0004
createHistorical Fiction	1.8089	0.3409	5.31	0.0000
createKids Fiction	1.5987	0.4655	3.43	0.0006
createScience Fiction	1.1859	0.3791	3.13	0.0018
createSuper Hero	1.7274	0.6198	2.79	0.0055
prod.dum1	1.1377	0.6902	1.65	0.0998
prod.dum2	0.4774	1.0833	0.44	0.6596
prod.dum3	1.3804	0.7257	1.90	0.0577
prod.dum4	1.0985	0.3867	2.84	0.0047
prod.dum6	1.0733	1.2725	0.84	0.3993
prod.dum7	1.2028	1.3794	0.87	0.3836
genre.dumAction/Adventure	0.4009	0.3613	1.11	0.2677
genre.dumAnimated	0.3542	1.1670	0.30	0.7616
genre.dumChildrens	0.4480	0.5886	0.76	0.4468
genre.dumComedies	0.9118	0.3441	2.65	0.0083
genre.dumDramas	0.5823	0.3436	1.69	0.0907
genre.dumEducation/General Interest	3.3366	0.7839	4.26	0.0000
genre.dumForeign Films	0.6764	0.5784	1.17	0.2427
genre.dumHorror/Suspense	1.4621	0.3734	3.92	0.0001
genre.dumRomance	1.0782	0.5653	1.91	0.0570
genre.dumScience-Fiction/Fantasy	1.3470	0.5379	2.50	0.0126
genre.dumThriller	0.6043	0.4885	1.24	0.2166
init.theat	0.0005	0.0003	1.72	0.0867
max.theat	0.0004	0.0003	1.08	0.2810
holiday1	0.2951	0.2541	1.16	0.2460
holiday2	0.0493	0.1986	0.25	0.8040
holiday3	-0.0935	0.2158	-0.43	0.6651
holiday4	0.1402	0.3034	0.46	0.6442
holiday5	-0.3917	0.4339	-0.90	0.3670
holiday6	-0.3671	0.2594	-1.42	0.1576
holiday7	-0.0483	0.1985	-0.24	0.8079
holiday8	-0.1441	0.2175	-0.66	0.5078
holiday9	-0.1840	0.4086	-0.45	0.6526
holiday10	0.0530	0.2255	0.23	0.8144

Table 8: Full coefficient table: OLS results for ‘independent’ movies (< 600 screens at maximum release). Dependent variable is profit (logged movie revenue minus logged movie costs).

References

- Ainslie, Andrew, Xavier Drèze and Fred Zufryden. 2005. "Modeling Movie Life Cycles and Market Share." *Marketing Science* 24(3):508–17.
- Blumer, Herbert. 1933. *Movies and Conduct*. New York: Macmillan.
- Bock, R and M Lieberman. 1970. "Fitting a response curve model for dichotomously scored items." *Psychometrika* 35:179–198.
- Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2).
- Eliashberg, Jehoshua and Steven M. Shugan. 1997. "Film Critics: Influencers or Predictors?" *Journal of Marketing* 61(2):68–78.
- Elsworthin, Catherine. 2005. "Sony to pay \$1.5m for film hoax." (*Dublin*) *Independent* August 5.
- Firth, David. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* 80:27–38.
- Hambleton, Ronald K., H. Swaminathan and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.
- Heckman, James. 1981. In *Structural Analysis of Discrete Data With Econometric Applications*, ed. C. Manski and D. McFadden. Cambridge, MA: MIT Press chapter The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence.
- Hollinger, Hy. 2007. "MPA study: Brighter Picture for Movie Industry." *Hollywood Reporter* June 15.
- Kracauer, Stanley. 1957. *From Caligari to Hitler: A Psychological History of the German Film*. Princeton, NJ: Princeton University Press.
- Laslier, Jean-Francois. 2005. "Spatial Approval Voting." *Political Analysis* 14(2):160–185.

- Lewis, Jeff and Keith Poole. 2004. "Measuring Bias and Uncertainty in Ideal Point Estimates via the Parametric Bootstrap." *Political Analysis* 12:105–127.
- Lord, Frederic M. 1980. *Application of Item Response Theory To Practical Testing Problems*. Mahwah NJ: Lawrence Erlbaum Associates.
- Martin, Andrew and Kevin Quinn. 2001. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis* 10(2).
- Mulvey, Laura. 1975. "Visual Pleasure and Narrative Cinema." *Screen* 16(3):6–18.
- Neelamegham, Ramya and Pradeep Chintagunta. 1999. "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets." *Marketing Science* 18(2):115–136.
- Poole, Keith and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35:228–278.
- Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political Economic History*. New York: Oxford University Press.
- Rasch, Georg. 1981. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Riesman, David, Revel Denny and Nathan Glazer. 1968. *The Lonely Crowd*. New Haven, CT: Yale University Press.
- Smith, Scott. 1998. *The Film 100: A Ranking of the Most Influential People in the History of the Movies*. Yucca Valley, CA: Citadel.
- Talbert, Jeffery C. and Matthew Potoski. 2002. "Setting the Legislative Agenda: The Dimensional Structure of Bill Cosponsoring and Floor Voting." *Journal of Politics* 64(3):864–891.
- van der Linden, Wim J and Ronald K. Hambleton. 1997. *Handbook of Modern Item Response Theory*. New York: Springer chapter Item Response Theory: Brief History, Common Models, and Extensions.