

On the analysis of coarsened normal data with unknown cut-points:
A comparison of the proportional-odds model with the normal theory *t*-test

Todd Bodner
Department of Psychology
Harvard University

May 11, 1999

Running Head: COARSENEDED NORMAL DATA

Abstract

Employing a situation analogous to a two group randomized experiment, a simulation experiment was derived to compare the results of the proportional-odds model and the normal theory *t*-test on data randomly drawn from normal distributions, but coarsened into ordinal categories under four different categorization schemes (i.e., 3 versus 7 outcome categories X "control group centered" cut-points versus "shifted-left" cut-points). In particular, each model's Type I error probability and statistical power was estimated under these categorization schemes with either a small sample ($n_t = n_c = 10$) or a large sample ($n_t = n_c = 50$). When large samples were employed, both models performed similarly across the categorization schemes. With the small samples, the proportional-odds model displayed better control of the Type I error probability and better statistical power, especially when 3 outcome categories were used.

On the analysis of coarsened normal data with unknown cut-points:

A comparison of the proportional-odds model with the normal theory *t*-test

Outcome data in the social and behavioral sciences often take the form of counts on an arbitrary integer-valued ordinal scale. Two statistical methods are typically invoked to analyze such data. The first method ignores the ordered relationship between the categories and analyzes the data using either an analysis of contingency tables, a log-linear statistical model, or a multinomial regression model. The second method assumes a latent normal distribution underlying the ordinal categories and uses normal theory statistical models to analyze the data (e.g., a *t*-test or *ANOVA*). The popularity of these methods is due in part to their relative simplicity and in part to the emphasis given them during a researcher's statistical training. However, statistical models to analyze ordinal data exist that do not require one to ignore the ordered nature of the outcome data nor to assume a normally distributed latent variable underlying the data. One such model is the proportional-odds model (e.g., McCullagh, 1980).

In the research literature, little has been written comparing these three types of models with one notable exception. There is a substantial literature suggesting that one should not ignore the ordinal nature of the outcome data. It is argued that such approaches ignore an important pattern in the data and lead to a necessary increase in the number of parameters needed to model the data

accurately (Agresti, 1990; Dobson, 1990; McCullagh & Nelder, 1989). However, I have found no explicit comparisons of statistical models for ordered category outcome variables with normal theory models. This seems to be a much more important comparison to be made than the comparison of models that ignore the category orderings with those models that do not ignore the category orderings. This paper is an initial attempt to fill this void in the research literature.

The proportional-odds model

In this paper, I assume familiarity with the normal theory t -test. However, practicing researchers may not be aware of the proportional-odds model. I briefly introduce this model here. For a more complete discussion, see McCullagh (1980) and McCullagh and Nelder (1989, Ch. 5). Rather than assuming an underlying continuous distribution like the t -test, the proportional-odds model seeks to estimate the logarithm of the odds of being in outcome category j or less relative to being in category $j+1$ or greater conditional on the covariates in the model for all j . This model is called proportional as the ratio of the cumulative odds under two possible values for a particular covariate,

$$P(Y \leq j \mid x = x_1) / P(Y \leq j \mid x = x_2),$$

depends only on the difference in the covariate values and not on the particular outcome category, j , under consideration. In this sense, differences in the cumulative odds are proportional to the differences in particular covariate values and are independent of the category classifications. Models do exist that do not require this proportionality, but they are more complicated and will not be

considered presently. In a simple model with one explanatory variable, a positive coefficient for a particular explanatory variable implies that higher values of the explanatory variable are associated with greater use of the higher outcome categories relative to lower values of the explanatory variable.

As the proportional-odds model does not rely on a particular underlying continuous distribution (e.g., a normal distribution), it would be interesting to compare the proportional-odds model with the normal theory *t*-test when the underlying distribution of the data is known. In this paper, I make this comparison in the special case in which the underlying distributions conditional on the covariates are normal.

Where are the cut-points in the normal distributions?

When data drawn from a normal distribution are coarsened, a natural question arises regarding location of the "cut-points" where this coarsening occurs. These categorization cut-points depend on many factors including the number of outcome categories employed and where these cut-points fall in relation to the underlying distributions.

The extent that the number and location of these cut-points influence the substantive conclusions of a statistical analysis is an open question. In the present study, I explore whether the number and location of these cut-points differentially affect the substantive conclusions of statistical analyses of coarsened normal data using the proportional-odds model and the *t*-test.

Simulation Study

This simulation study was designed to mimic a two-group experiment where the outcome data is ordinal in nature. In these simulations, the underlying treatment group distributions were normal, coarsened to consist of a various number of categories. The location of the cut-points in their relation to the underlying normal distribution was manipulated as was the sample sizes within each treatment group. The coarsened-normal data were then analyzed using the normal theory t -test and the proportional-odds model and their results are compared. Though rejecting or not rejecting a null hypothesis strictly on the basis of the p -values generated by any statistical test is a questionable research practice, it remains the single most influential quantity for assessing the worth of a statistical model by practicing researchers. As such, the primary outcome data in this simulation were the simulated probability of an $\alpha = 0.05$ Type I error and the statistical power¹ for each model under the various simulation conditions.

Design

Several simulation parameters were manipulated to explore for any differential sensitivity in the Type I error probability and statistical power between the normal-theory t -test and the proportional-odds model. These parameters can be categorized as parameters over which the researcher has some control and parameters over which the researcher has no control.

¹ Recall that a Type I error results when one rejects a null hypothesis when it is in fact true and statistical power is the probability of rejecting a false null hypothesis. For the remainder of this paper, when the nominal Type I error probability is set equal to 0.05.

Study design factors. Two of the simulation parameters are typically under the direct control of the researcher, the number of outcome categories employed and the size of the study. In the simulation, two quantities of outcome categories were used, 3 categories and 7 categories. When using 3 outcome categories, the integers running from 1 to 3 and when using 7 outcome categories, the integers running from 1 to 7 were used. Furthermore, two treatment group sample sizes were used (i.e., 10 and 50 subjects per treatment group). Both of these design factors cover a substantial portion of the ranges of outcome categories used and sample sizes used in experimental research in psychology.

Underlying distribution factors. Two factors that are not under the direct control of the researcher are the parameters of the underlying distributions and the location of the cut-points in relation to the underlying distributions that coarsen the data into the ordinal outcome categories. For the simulation, the underlying treatment group distributions were normal with variance equal to one. For exploring potential differences in Type I error probabilities under each model, both treatment means were set equal to zero. For exploring potential differences in sensitivity to statistical power under each model, the control² group mean was set equal to zero and the treatment group mean was set equal to 0.5. This difference of 0.5 between the treatment and control groups was chosen

² For communication purposes, one of the treatment groups is referred to as the "control" group.

as it coincides with a convention in behavioral research for a treatment effect that is *moderate* in size (Cohen, 1977).

The locations of the cut-points were assigned to create equal-length intervals in the range from -3 to 3. For a standard normal distribution, the interval [-3,3] constitutes most of the probability density. The number of these intervals of course depended on the number of outcome categories used. The locations of the cut-points were based on the quantiles of a uniform [-3,3] necessary to achieve i equal intervals, where i is an index for the number of outcome categories.

The locations of these cut-points in relation to the underlying normal distributions were manipulated in the following fashion. The cut-points locations were always defined in relation to the underlying control group distribution. When the cut-points were the quantiles of the control group distribution necessary to achieve equal intervals in the range [-3,3], this corresponds to a shift of "0" and will be referred to as a "centered" location of the cut-points. In contrast with the centered cut-points, there was a "shift-left" condition in which the centered cut-points were shifted to lower values by 1. In this simulation, the same cut-points coarsened both the treatment and control group distributions. Figure 1 below gives an example of how the outcome category placement was achieved for the case in which there was a treatment effect and the cut-points were centered for the underlying control group distribution.

Figure 1. Example of the location of the "centered" and "left-shifted" cut-point locations when 3 outcome categories were used in the presence of a treatment effect.

Figure not available . . . Sorry.

The two numbers of outcome categories considered crossed with the two possible shifting parameters yielded four possible combinations of cut-point placement and the number of outcome categories. These four combinations are collectively referred to as the "categorization schemes" though this is not meant to imply that all of these factors are controllable by the researcher.

The probabilities of the random data being categorized into the outcome categories under the various categorization schemes are presented below for the case in which there is no treatment effect (see Table 1) and when there is a moderate treatment effect (see Table 2). It should be noted that the categorization probabilities are the same for the control group whether or not there is a treatment effect.

Table 1. Category probabilities for the treatment groups with no treatment effect for the four categorization schemes.

	Three Outcome Categories			Seven Outcome Categories						
	1	2	3	1	2	3	4	5	6	7
Centered Cut-Points	0.16	0.68	0.16	0.02	0.08	0.23	0.33	0.23	0.08	0.02
Left-Shifted Cut-Points	0.02	0.48	0.50	0.001	0.01	0.07	0.21	0.33	0.26	0.13

Table 2. Category probabilities for the *treatment* group under the condition of a $+0.5$ mean difference with the control group for the four categorization schemes.

	Three Outcome Categories			Seven Outcome Categories						
	1	2	3	1	2	3	4	5	6	7
Centered Cut-Points	0.07	0.63	0.30	0.004	0.03	0.14	0.30	0.31	0.17	0.05
Left-Shifted Cut-Points	0.01	0.30	0.69	0.0001	0.003	0.02	0.12	0.27	0.32	0.26

Procedure

Two miniature simulations were conducted in this study, one corresponding to a within-group sample size of 10 and one corresponding to a within-group sample size of 50. An annotated copy of the simulation program is attached in Appendix A. Within each step of each miniature simulation the following procedure was used. A standard normal random variable was drawn for each of the "subjects" in each group. For the simulation conditions in which the treatment group mean differed from the control group mean, 0.5 was added to each random standard normal variable drawn in the treatment condition. These random values were then categorized into the integer-valued outcome categories (i.e., 3 or 7 outcome categories) under the two shifts of the cut-points (i.e., -1 for the left-shift conditions and 0 for the centered conditions).

Two statistical tests were then conducted on these coarsened data, an independent two-group *t*-test assuming equal variances and the proportional-odds model³. The *p*-values under each model for each of the three and seven category outcome data were then recorded under the two cut-point shift conditions.

The following methods were used in some of the troublesome situations that occurred in the small sample case with few outcome categories. When there was no variation in the outcome data for either treatment group and the

treatment group means were identical, the p -value of 1.00 was assigned. When there was no variation in the outcome data within each treatment group and the treatment group means were not identical, a Fisher's exact test was conducted and that p -value was recorded for the two tests. The number of such exact tests and $p = 1.00$ assignments that were required was recorded.

Thus, each of the two simulations yielded 8 sets of counts (i.e., 2 (number of outcome categories) X 2 (shift cut-points left or centered) X 2 (same distribution means or different distribution means)). Within each set were four numbers, the number of exact tests and $p = 1$ assignments required and the computed p -values under the two models. Each simulation was iterated 2000 times.

Simulation Results

The discussion of the simulation results is presented by first discussing the differences in the statistical models for the large sample size followed by a discussion of the results for the small sample size.

Large Sample Sizes

In none of the conditions under the large sample size was an exact test or the assignment of a p -value of 1.00 required. For large samples, the proportional-odds model and the t -test give similar results in terms of the probability of Type I error occurrence and statistical power. These details are briefly presented.

³ The proportional-odds model was computed on S-plus from the "lrm" function available within a library written by Frank Harrell that is available in the "Design" library on the Carnegie Mellon

Probability of a Type I error. The following results were estimated in the conditions in which the underlying treatment and control distributions were assigned the same mean. Estimated p -values less than 0.05 for each model on the simulated data constituted a Type I error. The number and proportion of such Type I errors for each model under the four categorization schemes were recorded. As the following table depicts, the proportions of Type I errors committed under each categorization scheme for each model were relatively constant across the conditions in the simulation and were very close to the criterion significance level of 0.05.

Table 3. The number and proportion of $\alpha = 0.05$ Type I errors committed in which there was no difference between the treatment and control distribution means ($n_t = n_c = 50$) in 2000 simulation runs.

		Number of Outcome Categories				Margin	
		Three		Seven		<i>t</i> -test	<i>POM</i>
		<i>t</i> -test	<i>POM</i>	<i>t</i> -test	<i>POM</i>		
Location of Cut-points	Centered	110 0.055	110 0.055	106 0.053	104 0.052	216 0.054	214 0.054
	Left	109 0.055	102 0.051	105 0.053	110 0.055	214 0.054	212 0.053
	Margin	219 0.055	212 0.053	211 0.053	214 0.054	430 0.054	426 0.053

Note: The first number in each cell is the number of p -values less than 0.05 out of 2000 and the second number is the proportion of p -values less than 0.05 out of 2000.

There are no substantial differences between the two models in their estimated probabilities of a Type I error under the various categorization schemes.

Statistical power. The statistical power of each model under the various categorization schemes was explored by creating a difference between the treatment and control conditions by adding 0.5 to each data value in the treatment group. Of interest was any differential ability for each model to detect this difference in the underlying distributions under the various categorization schemes. The number and proportion of correct rejections of the null hypothesis at the $\alpha = 0.05$ significance level were recorded and compared.

The number and proportion of correct rejections of the false null hypothesis are very close for the *t*-test and the proportional-odds model in each categorization condition of the simulation as the following table depicts.

Table 4. The number and proportion of $\alpha = 0.05$ correct rejections of a false null hypothesis in the 2000 simulation runs in which there was a moderate effect-size difference between the treatment and control distribution means ($n_t = n_c = 50$).⁴

		Number of Outcome Categories				Margin	
		Three		Seven		<i>t</i> -test	<i>POM</i>
		<i>t</i> -test	<i>POM</i>	<i>t</i> -test	<i>POM</i>		
Location of Cut-points	Centered	1147 0.574	1152 0.576	1336 0.668	1316 0.658	2483 0.621	2468 0.617
	Left	1033 0.517	1026 0.513	1334 0.667	1317 0.659	2367 0.592	2343 0.586
	Margin	2180 0.545	2178 0.545	2670 0.668	2633 0.658	4850 0.606	4811 0.601

With the large sample size and seven outcome categories, the shifting of the category cut-points had no noticeable effect on the power of the statistical tests. As a result, differences in the models will not be compared.

However in lieu of this discussion, the effects of categorization scheme on statistical power of the models will be compared. For the characteristics of the underlying normal distributions of interest, the nominal statistical power to detect a difference in means is approximately 0.71. When seven outcome categories are used, the estimated statistical power of these models approaches the expected power under the normal model. Marginally, there appears to be an approximate drop in power of 17% when one uses three outcome categories rather than seven. This drop is approximately twice as large for non-centered cut-points (i.e., ~ 27%) compared to centered cut-points (i.e., ~ 14%). Marginally, there appears to be a small decrease in power when the cut-points are not

⁴ See the note to Table 3 for a definition of the quantities in each of the cells.

centered on the control group distribution (i.e., ~ 4.1%). However, this drop appears to be larger when three outcome categories are used (i.e., ~ 9.9%) compared to when seven outcome categories are used (i.e., ~ 0.07%).

Summary. When sample sizes are large, both the proportional-odds model and the t -test perform similarly with respect to their Type I error probabilities as well as their statistical power.

Small Sample Sizes

The estimates of differences in the Type I error probabilities and the statistical power between the proportional-odds model and the t -test in the small sample case is more complex and interesting. Differences in these estimates between the two models are found and these differences appear to be in part a function of the categorization scheme in effect.

Probability of a Type I error. In the simulations for estimating the Type I error probabilities under small sample conditions, two of the four categorization conditions created data in which an exact test or assignment of a p -value of 1.00 was required. Both of these conditions involved the use of three outcome categories. When the location of the cut-points were centered, 71 exact tests and 9 $p = 1.00$ assignments were necessary. When the location of the cut-points were shifted left, 3 exact tests were necessary.

The number and proportion of p -values less than the nominal $\alpha = 0.05$ vary slightly depending on how the continuous data were categorized and which

statistical model was used to analyze the ordinal categorical data as can be seen in Table 5 below.

Table 5. The number and proportion of p -values less than $\alpha = 0.05$ in the 2000 simulation runs in which there was no difference between the treatment and control distribution means ($n_t = n_c = 10$).⁴

		Number of Outcome Categories				Margin	
		Three		Seven			
		<i>t</i> -test	<i>POM</i>	<i>t</i> -test	<i>POM</i>	<i>t</i> -test	<i>POM</i>
Location of Cut-points	Centered	91 0.046	152 0.076	89 0.045	107 0.054	180 0.045	259 0.065
	Left	79 0.040	108 0.054	83 0.042	113 0.057	162 0.041	221 0.055
	Margin	170 0.043	260 0.065	172 0.043	220 0.055	342 0.043	480 0.060

Overall, the *t*-test committed slightly more $\alpha = 0.05$ Type I errors and the proportional-odds model committed slightly fewer $\alpha = 0.05$ Type I errors than the nominal 5% would dictate though these differences are quite small. To ease a comparison of the models, the data within each categorization scheme were combined into a single metric, the increase or decrease in the number or proportion of p -values less than $\alpha = 0.05$ for the proportional-odds model relative to the *t*-test. The following table displays these values.

Table 6. Relative change in the proportional-odds model relative to the *t*-test in the obtained number of p -values less than $\alpha = 0.05$ when there was no difference between the treatment and control distribution means ($n_t = n_c = 10$).

		Number of Outcome Categories		Margin
		Three	Seven	
Location of Cut-points	Centered	-67.03%	-20.22%	-43.89%
	Left	-36.71%	-36.14%	-36.42%
	Margin	-52.94%	-27.91%	-40.35%

In each condition in the simulation with a small sample size, the proportional-odds model committed fewer $\alpha = 0.05$ Type I errors. Overall, the proportional-odds model was approximately 40% less likely to produce an $\alpha = 0.05$ Type I error relative to the t -test in this simulation.

Marginally, the number of outcome categories had a moderate effect on the relative number of Type I errors committed by each model. Using seven rather than three outcome categories lead to a reduction in the relative number of Type I errors committed by each model (i.e., -52.94% compared to -27.91%). This marginal difference is almost entirely due to the decrease in the relative number of Type I errors under each model when using seven outcome categories (i.e., -20.22%) compared to three outcome categories (i.e., -67.03%) when the categorization cut-points were centered on the control group distribution. This reduction is almost non-existent when the cut-points are shifted lower in the control group distribution (i.e., -36.71% for three outcome categories versus -36.14% for seven outcome categories).

Marginally, there is a small effect of the location of categorization cut-points in the control group distribution on the relative number of Type I errors committed by each model. The proportional-odds model was approximately 44% less likely to commit a Type I error relative to the t -test when the cut-points were centered in the control group distribution and approximately 36% less likely to commit a Type I error relative to the t -test when the cut-points were

shifted lower in the control group distribution. However, this marginal difference in the relative number of Type I errors between the centered and left-shifted cut-points was only apparent in the three outcome category conditions. In these conditions, the proportional-odds model was approximately 67% less likely to commit a Type I error relative to the t -test when the cut-points were centered in the control group distribution and approximately 37% less likely to commit a Type I error relative to the t -test when the cut-points were shifted lower in the control group distribution. For the seven outcome category conditions, the direction of this difference reversed and was smaller. In particular for the seven outcome category conditions, the proportional-odds model was approximately 20% less likely to commit a Type I error relative to the t -test when the cut-points were centered in the control group distribution and approximately 36% less likely to commit a Type I error relative to the t -test when the cut-points were shifted lower in the control group distribution.

Statistical power. In the simulation to estimate the statistical power of each model under the various categorization schemes, the same two conditions as mentioned above required the use of an exact test or the assignment of a p -value of 1.00. Again, both conditions involved the use of three outcome categories. When the cut-points were centered, 52 exact tests and 6 $p = 1.00$ assignments were necessary. When the cut-points were shifted left, 53 exact tests were necessary.

The statistical power of both the proportional-odds model and the t -test were found to differ at times depending on the categorization scheme as Table 7 below shows.

Table 7. The number and proportion of $\alpha = 0.05$ correct rejections of a false null hypothesis in which there was a moderate effect-size difference between the treatment and control distribution means ($n_t = n_c = 10$) in 2000 simulation runs.⁴

		Number of Outcome Categories				Margin	
		Three		Seven			
		<i>t</i> -test	<i>POM</i>	<i>t</i> -test	<i>POM</i>	<i>t</i> -test	<i>POM</i>
Location of Cut-points	Centered	305 0.153	426 0.213	365 0.183	415 0.208	670 0.168	841 0.210
	Left	212 0.106	275 0.138	352 0.176	393 0.197	564 0.141	668 0.167
	Margin	517 0.129	701 0.175	717 0.179	808 0.202	1234 0.154	1509 0.189

The nominal statistical power based on the characteristics of the underlying normal distributions is approximately 0.20 assuming $\alpha = 0.05$. With such small sample sizes, this is a low statistical power situation though the effective true difference in the treatment and control distributions is moderate in size. In every condition of the simulation, the proportional-odds model exhibited greater statistical power than the *t*-test.

Again, to aid in the differential effect of the categorization schemes on the statistical power for each model, the relative difference in statistical power of the proportional-odds model relative to the *t*-test was used. These quantities are presented in the table below.

Table 8. Relative change in the statistical power of the proportional-odds model over the *t*-test where there was a moderate effect size difference between the treatment and control distribution means ($n_t = n_c = 10$).

		Number of Outcome Categories		
		Three	Seven	Margin
Location of	Centered	+39.67%	+13.69%	+25.52%

Cut-points	Left	+29.72%	+11.65%	+18.44%
	Margin	+35.59%	+12.69%	+22.29%

Over all of the categorization scheme conditions, the proportional-odds model was approximately 22% more powerful than the *t*-test. Marginally, this proportional difference in power was greater for the three outcome category conditions (i.e., +35.59%) compared to the seven outcome category conditions (i.e., +12.69%). This was especially true in those conditions in which the categorization cut-points were centered in the control group distribution where the proportional difference in power was greater for the three outcome categories (i.e., 39.67%) than for the seven outcome categories (i.e., 13.69%). This difference was slightly smaller when the categorization cut-points were shifted lower in the control group distribution where the proportional difference in power was greater for the three outcome categories (i.e., 29.72%) than for the seven outcome categories (i.e., 11.65%).

Marginally, there was a small effect for the location of the categorization cut-points on the relative statistical power under each model. The relative increase in power of the proportional-odds model over the *t*-test was greater in when the categorization cut-points were centered (i.e., 25.52%) compared to when they were shifted lower (i.e., 18.44%). This difference was especially true when three outcome categories were used where the relative increase in power of the proportional-odds model over the *t*-test was greater when the categorization

cut-points were centered (i.e., 39.67%) compared to when they were shifted lower (i.e., 29.72%). This difference was smaller when seven outcome categories were used (i.e., 13.69% for centered cut-points versus 11.65% for left-shifted cut-points).

Summary. Overall, it appears that when sample sizes are small, use of the proportional-odds model over the *t*-test gives two benefits. First, it seems to afford greater protection against Type I errors and second it seems to offer greater statistical power.

Discussion

The operating characteristics of the proportional-odds model and the *t*-test were compared in the analysis of coarsened-normal data under various forms of categorization. In particular, the relative effects of the number of outcome categories used by the researcher and the location of the categorization cut-points in the underlying distributions on the Type I error probabilities and statistical power for each model were contrasted.

It was found that with large sample sizes, no important differences were found between the two models; both models show appropriate and similar control of the probability of a Type I error and similar levels of statistical power. With small sample sizes, it was found that the proportional-odds model offered better protection against the probability of the Type I error as well as greater statistical power.

Based on the results of this simulation, a few tentative recommendations can be made though more extensive work on this subject is needed to clarify and verify these results. It appears that one should try to have a larger rather than smaller set of ordinal outcome categories to best preserve statistical power and counteract the effects of unknown locations of categorization cut-points on the underlying distributions. In such cases, the differences between the results of an analysis under the proportional-odds model and the t -test will tend to offer similar control over Type I error probability as well as comparable statistical power when the sample sizes are large. However, when the number of outcome categories is small rather than large and the sample sizes are small, the proportional-odds model appears to offer greater protection against the Type I error and greater statistical power across the location of unknown categorization cut-points in the underlying distributions over the t -test.

Qualifications, questions, and recommendations for future work. The stated results are tentative as they are critically dependent on two important factors, the absence of errors in the simulation program and an unknown degree of variability in the estimates found. Regarding the first factor, there might very well be an error in the program used to compare these models. An annotated copy of the program is included in Appendix A. Regarding the second factor, there is certainly a fair amount of variability in the quantities estimated. Assuming the program is working correctly, this variability could be better

controlled by increasing the number of iterations in the simulation beyond the 2000 used in the present study.

One aspect of the results that deserves further attention is the finding that in terms of statistical power, the two models tended have less of a reduction in statistical power in relation to the nominal power when the sample sizes were small compared to when the sample sizes were large. This is very likely due to simulation uncertainty in the small sample case, but this should be checked.

It is important to note that these results are only relevant to studies in which the underlying treatment and control distributions are normal with common variance with equal interval categorization cut-points. Future work should consider when the underlying distributions are normal without a common variance or non-normal and with categorization cut-points that are other than the equal interval cut-points employed herein. One might speculate that the relative similarity of the proportional-odds model and the t -test in this study is in part due to the fact that the underlying distributions *are* normal. It is interesting that the proportional-odds model does so well in this case relative to the t -test. One might speculate that the proportional-odds model would prove to be even more useful than the t -test when the underlying distributions are far from normality.

In conclusion, it appears that both the proportional-odds model and the t -test do an adequate and in many cases a comparable job analyzing coarsened-normal data. While the t -test makes the assumption of an underlying normal

distribution, the proportional-odds model does not. As a result, using the proportional-odds model might be useful when an underlying normal distribution is unknown or cannot be assumed.

References

- Agresti, A. (1990). Categorical data analysis. New York: John Wiley.
- Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.
- Dobson, A. (1990). An introduction to generalized linear models. New York: Chapman & Hall/CRC Press.
- McCullagh, P. (1980). Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society, Series B, 42, 109-142.
- McCullagh, P., & Nelder, J. (1989). Generalized linear models (2nd ed.). New York: Chapman & Hall.

Appendix A

Annotated S-Plus Simulation Program

```

> simulate <- function(n, iters, teffect, shift, cats)
+ {
+   ### These commands set up the vectors to hold particular
+   ### results.
+   pvalues <- matrix(0, nrow=iters, ncol=2)
+   Ytf <- c(1:n)
+   Ycf <- c(1:n)
+   Xt <- rep(1, n)
+   Xc <- rep(0, n)
+   X <- append(Xt, Xc)
+   numexact <- 0
+   assign1 <- 0
+   ### These commands create the shifts in the quantile cut-points
+   sh <- rep(shift, (cats - 1))
+   if(cats==3)
+   {
+   sc <- threecut + sh
+   }
+   if(cats==7)
+   {
+   sc <- sevenscut + sh
+   }
+   for(iter in 1:iters)
+   {
+   if((iter/iters)==(2|4|6|8|10))
+   {cat("iter:", "", iter)}
+   ### The following commands extract randomly generated
+   ### standard normal random data from a previously defined
+   ### matrix. If a treatment effect is non-zero, this
+   ### value is added to the treatment data.
+   Ytf <- Y[iter, (n+1):(2*n)] + teffect
+   Ycf <- Y[iter, 1:n]
+   ### The following commands categorize the random normal
+   ### data into the outcome categories
+   if(cats==3)
+   {
+   Yt <- ifelse(Ytf<=sc[1], 1,
+               ifelse(Ytf<=sc[2], 2, 3))
+   Yc <- ifelse(Ycf<=sc[1], 1,
+               ifelse(Ycf<=sc[2], 2, 3))
+   }
+   if(cats==7)
+   {
+   Yt <- ifelse (Ytf>sc[6], 7,
+               ifelse(Ytf>sc[5], 6,
+               ifelse(Ytf>sc[4], 5,
+               ifelse(Ytf>sc[3], 4,
+               ifelse(Ytf>sc[2], 3,
+               ifelse(Ytf>sc[1], 2, 1))))))
+   Yc <- ifelse (Ycf>sc[6], 7,

```

```

+             ifelse(Ycf>sc[5], 6,
+             ifelse(Ycf>sc[4], 5,
+             ifelse(Ycf>sc[3],4,
+             ifelse(Ycf>sc[2],3,
+             ifelse(Ycf>sc[1],2,1))))))
+         }
+     y <- append(Yt, Yc)
+     ### The next set of commands conduct the statistical
+     ### analyses on the categorized data.
+     ### This first "if" commands checks to make sure that
+     ### the treatment and control distributions have
+     ### sufficient variability between the groups and within
+     ### the groups for the statistical analyses.
+     if( (abs(sum(Yt)-sum(Yc))<1) && ((var(Yt) < 0.02) ||
+     (var(Yc) < 0.02)))
+     {
+         ### These are the p-value = 1 assignments under
+         ### insufficient within and between group
+         ### variability. Their values are set equal to
+         ### 2 so that they could be identified easier in
+         ### the output.
+         tep <- 2
+         pomp <- 2
+         assign1 <- assign1 + 1
+     }
+     else
+     {
+         ### These "if" commands check to see if there is
+         ### sufficient within group variability to conduct
+         ### the tests. If there is, the standard tests are
+         ### conducted. If not, a Fisher exact test is
+         ### is conducted.
+         if ((var(Yt) > 0.02) && (var(Yc) > 0.02))
+         {
+             te <- t.test(Yt,Yc)
+             tep <- te$p.value
+             pom <- lrm(y ~ X)
+             pomp <- pom$stats[5]
+         }
+         if ((var(Yt) <= 0.02) || (var(Yc) <= 0.02))
+         {
+             exact <- fisher.test(table(y,X))
+             tep <- exact$p.value
+             pomp <- exact$p.value
+             numexact <- numexact + 1
+         }
+     }
+     ### Combining the obtained p-values
+     p <- c(tep,pomp)
+     pvalues[iter, ] <- p
+ }
+ list(pvalues=round(pvalues,6),numexact=numexact,assign1=assign1)
+ }

```

Issues to consider

Statistical power versus precision of the estimate (see Maxwell, 1998, PM, 279)

How do the null hypotheses under each model compare?