Example 3a: Generalized Linear Models for Binomial Outcomes (% Correct) using SAS and STATA
(complete syntax and output available for SAS and STATA electronically)

The data for this example come from the publication below, which examined annual growth in a test of grammatical understanding from Kindergarten through 4th grade in children with non-specific language impairment (NLI) or specific language impairment (SLI):

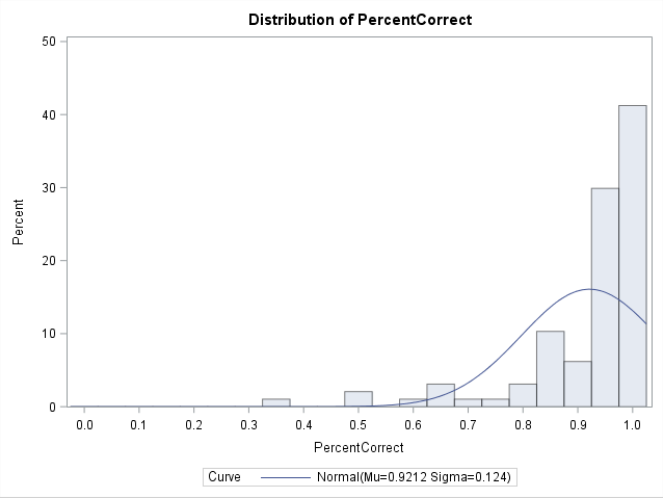
Rice, M. L., Tomblin, J. B., **Hoffman, L.**, Richman, W. A., & Marquis, J. (2004). Grammatical tense deficits in children with SLI and nonspecific language impairment: Relationships with nonverbal IQ over time. *Journal of Speech-Language-Hearing Research, 47*(4), 816-834.

The current example is a cross-sectional analysis of how grammatical understanding at third grade is predicted by group (NLI=0, SLI=1) and mother's years of education (centered so that 0=12 years). Given that percent correct is bounded by 0 and 1, we will use a logit link and the binomial family of conditional response distributions. Because the binomial is a discrete distribution, we will need to parameterize the model to predict the number of correct responses out of the number of trials instead of percent correct. This example will also demonstrate two ways of addressing binomial overdispersion: *additive* (through individual random intercepts) and *multiplicative* (through the beta-binomial distribution), as well as zero-inflated (actually one-inflated here; stay tuned) versions of the binomial and beta-binomial model variants. In SAS GLIMMIX I am not using denominator DF so that the results match those of STATA as closely as possible.

SAS Data Manipulation and Description:

```
* Create predictor variables;
DATA work.Example3a; SET work.growthdata;
     IF class=2 THEN NLIvsSLI=0; * NLI;
     IF class=3 THEN NLIvsSLI=1; * SLI;
     momed12=mom_edc-12; * Mom ed (0=12);
* Subset to wave 4 (third grade) and complete cases;
     WHERE index1=4 AND NMISS(Ncorrect,NLIvsSLI,momed12)=0;
* Create number correct for binomial model;
     Ntrials=100; PercentCorrect=CompTns;
     Ncorrect=ROUND(PercentCorrect*Ntrials,1);
* Compute number incorrect for zero-inflated binomial model;
     Nincorrect=Ntrials-Ncorrect;
RUN;
* Export data to STATA format;
PROC EXPORT DATA=work.Example3a OUTFILE="&filesave.\Example3a.dta" DBMS=STATA REPLACE; RUN;

TITLE "Distribution of Percent Correct";
PROC UNIVARIATE NOPRINT DATA=work.Example3a; VAR PercentCorrect;
     HISTOGRAM PercentCorrect / MIDPOINTS= 0 TO 1 BY .05 NORMAL(MU=EST SIGMA=EST); RUN; QUIT;
PROC MEANS NDEC=3 DATA=work.Example3a; VAR PercentCorrect; RUN; TITLE;
```



Individual mean % correct across 97 persons:
M=.9212, SD=.1240, Min=.3548, Max=1.00

Even though our distributional assumptions will be about the conditional outcome, not the original outcome, odds aren't good it will be normal!

But it may not be strictly binomial, either. The long tail to the left indicates possible over-dispersion (i.e., more variance leftover than the binomial distribution would predict), and there may be too many one values. We'll need to use models to test these suspicions empirically…

STATA Data Manipulation and Description:

```
* Import data
use "$filesave\Example3a.dta", clear
* Distribution of Percent Correct
hist percentcorrect, percent start(0) width(.05)

* Find and install betabin (and zbin and zibbin)
search betabin // install before continuing
```

---

1) Empty Means Binomial Model for % correct using DV = Events/Trials

$\#Correct_i \sim Binomial(p_i, Ntrials_i)$ → p_i is probability of any one trial being correct

$Logit(p_i \text{ for correct trial}) = \beta_0$

Conditional mean for $\#Correct_i = Ntrials_i * p_i$

Conditional variance for $\#Correct_i = (Ntrials_i * p_i)(1 - p_i)$

```
display as result "STATA Empty Means Binomial Model"
glm ncorrect, link(logit) family(binomial ntrials),
estat ic, n(97),
nlcom 1/(1+exp(-1*(_b[_cons]))) // intercept in probability (ILINK)


TITLE "SAS Empty Means Binomial Model";
PROC GLIMMIX DATA=work.Example3a NOCLPRINT NAMELEN=100 GRADIENT;
     MODEL Ncorrect/Ntrials = / SOLUTION DDFM=NONE LINK=LOGIT DIST=BINOMIAL;
     ESTIMATE "Intercept" intercept 1 / ILINK; * ILINK gives intercept in probability;
RUN; TITLE;
```

STATA Output:

```
Generalized linear models                         No. of obs      =         97
Optimization     : ML                             Residual df     =         96
                                                  Scale parameter =          1
Deviance         =   1620.05009                   (1/df) Deviance =   16.87552
Pearson          =  2041.435988                   (1/df) Pearson  =   21.26496
Variance function: V(u) = u*(1-u/ntrials)         [Binomial]
Link function    : g(u) = ln(u/(ntrials-u))       [Logit]
                                                  AIC             =   19.00196
Log likelihood   = -920.5951086                   BIC             =   1180.878
------------------------------------------------------------------------------
             |                 OIM
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   2.459276   .0376936    65.24   0.000     2.385397    2.533154
------------------------------------------------------------------------------
Akaike's information criterion and Bayesian information criterion

-------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+-----------------------------------------------------------------
           . |         97         .  -920.5951       1     1843.19   1845.765
-------------------------------------------------------------------------------
. nlcom 1/(1+exp(-1*(_b[_cons]))) // intercept in probability (ILINK)
------------------------------------------------------------------------------
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .9212371    .002735   336.83   0.000     .9158766    .9265977
------------------------------------------------------------------------------
```

SAS Output:

```
          Fit Statistics
-2 Log Likelihood            1841.19
AIC  (smaller is better)     1843.19
AICC (smaller is better)     1843.23
BIC  (smaller is better)     1845.76
CAIC (smaller is better)     1846.76
HQIC (smaller is better)     1844.23
Pearson Chi-Square           2041.44
Pearson Chi-Square / DF        21.26
```

To inverse link from logits to predicted % correct:

$$\text{Prob}(y = 1) = \frac{\exp(2.4593)}{1 + \exp(2.4593)} = .9212$$

The sample average probability of getting each item correct is .9212.

But Chi-Square/DF > 1, indicating that this model has over-dispersion (too much variance, likely in part because we haven't incorporated predictors yet).

```
                        Parameter Estimates
                     Standard
Effect       Estimate       Error       DF    t Value    Pr > |t|    Gradient
Intercept      2.4593     0.03769    Infty      65.24      <.0001    -1.03E-6

                                  Estimates
                     Standard                                              Standard Error
Label        Estimate       Error       DF    t Value    Pr > |t|        Mean        Mean
Intercept      2.4593     0.03769    Infty      65.24      <.0001      0.9212    0.002735
```

So even though we are modeling number of correct trials as the DV, the model is phrased to predict percent correct directly (as the conditional mean *p*, the probability that any trial = 1).

---

2) Two-Predictor Binomial Model

$\#Correct_i \sim Binomial(p_i, Ntrials_i)$ → p_i is probability of any one trial being correct

$Logit(p_i \text{ for correct trial}) = \beta_0 + \beta_1(NLIvsSLI_i) + \beta_2(MotherEd_i - 12)$

Conditional mean: $\#Correct_i = Ntrials_i * p_i$

Conditional variance of $\#Correct_i = (Ntrials_i * p_i)(1 - p_i)$

```
display as result "STATA Two-Predictor Binomial Model"
glm ncorrect c.nlivssli c.momed12, link(logit) family(binomial ntrials),
estat ic, n(97),
test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model

TITLE "SAS Two-Predictor Binomial Model";
PROC GLIMMIX DATA=work.Example3a NOCLPRINT NAMELEN=100 GRADIENT;
     MODEL Ncorrect/Ntrials = NLIvsSLI momed12
           / SOLUTION DDFM=NONE LINK=LOGIT DIST=BINOMIAL ODDSRATIO(LABEL);
     CONTRAST "Multiv Wald test of Model" NLIvsSLI 1, momed12 1 / CHISQ;
RUN; TITLE;
```

STATA Output:

```
Generalized linear models                         No. of obs      =         97
Optimization     : ML                             Residual df     =         94
                                                  Scale parameter =          1
Deviance         =  1310.593044                   (1/df) Deviance =   13.94248
Pearson          =  1448.891028                   (1/df) Pearson  =   15.41373
Variance function: V(u) = u*(1-u/ntrials)         [Binomial]
Link function    : g(u) = ln(u/(ntrials-u))       [Logit]
                                                  AIC             =   15.85292
Log likelihood   = -765.8665858                   BIC             =   880.5702
------------------------------------------------------------------------------
             |                 OIM
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nlivssli |  -1.221578   .0858707   -14.23   0.000    -1.389881   -1.053275
     momed12 |   .1193325   .0214268     5.57   0.000     .0773368    .1613283
       _cons |   3.071929   .0746183    41.17   0.000      2.92568    3.218178
------------------------------------------------------------------------------
```

```
Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |         97         .  -765.8666       3    1537.733   1545.457
-----------------------------------------------------------------------------
. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =  273.58
         Prob > chi2 =    0.0000
```

SAS Output:

```
           Fit Statistics
-2 Log Likelihood            1531.73
AIC  (smaller is better)     1537.73
AICC (smaller is better)     1537.99
BIC  (smaller is better)     1545.46
CAIC (smaller is better)     1548.46
HQIC (smaller is better)     1540.86
Pearson Chi-Square           1448.89
Pearson Chi-Square / DF        15.41 → better, but nowhere good enough!

                            Parameter Estimates
                         Standard
Effect       Estimate       Error       DF    t Value    Pr > |t|    Gradient
Intercept      3.0719     0.07462    Infty      41.17      <.0001    -5.77E-6
NLIvsSLI      -1.2216     0.08587    Infty     -14.23      <.0001    -3.06E-9
momed12        0.1193     0.02143    Infty       5.57      <.0001    -6.69E-6

                            Odds Ratio Estimates
                                                        95% Confidence
Comparison                              Estimate       DF          Limits
unit change of NLIvsSLI from mean          0.295    Infty       0.249       0.349
unit change of momed12 from mean           1.127    Infty       1.080       1.175

                                        Contrasts
Label                          Num DF   Den DF    Chi-Square    F Value      Pr > ChiSq    Pr > F
Multiv Wald test of Model           2    Infty        273.58     136.79          <.0001    <.0001
```

Before interpreting these results, let's see if we can get better distribution fit. Here are some alternative models that incorporate either overdispersion, zero-inflation (actually one-inflation here), or both…

---

3) Two-Predictor Binomial Model with <u>Additive Over-Dispersion</u>

$\#Correct_i \sim Binomial(p_i, Ntrials_i)$ → p_i is probability of any one trial being correct

$Logit(p_i \text{ for correct}) = \beta_0 + \beta_1(NLIvsSLI_i) + \beta_2(MotherEd_i - 12) + e_i$

Conditional mean of $\#Correct_i = Ntrials_i * p_i$

Conditional variance of $\#Correct_i = (Ntrials_i * p_i)(1 - p_i)$

The residual variance σ_e^2 is on the model-scale (in logits), and it effectively soaks up all discrepancy to each individual's predicted logit.

```
display as result "STATA Previous Two-Predictor Binomial Model"
display as result "Switch to MEGLM to do LRT against next model"
meglm ncorrect c.nlivssli c.momed12, link(logit) family(binomial ntrials),
estimates store FitBin  // save fit stats for binomial baseline

display as result "STATA Two-Predictor Binomial Model with Additive Overdispersion"
meglm ncorrect c.nlivssli c.momed12, || id: , /// || id. adds "residual variance"
      link(logit) family(binomial ntrials) intmethod(laplace),
estat ic, n(97),
test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
estimates store FitAOD  // save fit stats for model to compare
lrtest FitAOD FitBin    // LRT for additive overdispersion
```

```
display as result "STATA Two-Predictor Binomial Model with Additive Overdispersion"
display as result "Get Odds Ratios"
meglm ncorrect c.nlivssli c.momed12, || id: , ///  || id. adds "residual variance"
      link(logit) family(binomial ntrials) intmethod(laplace) eform,

TITLE "SAS Two-Predictor Binomial Model with Additive Overdispersion";
PROC GLIMMIX DATA=work.Example3a NOCLPRINT NAMELEN=100 METHOD=LAPLACE GRADIENT;
     CLASS ID; * Person ID is added to CLASS because of RANDOM statement below;
     MODEL Ncorrect/Ntrials = NLIvsSLI momed12
             / SOLUTION DDFM=NONE LINK=LOGIT DIST=BINOMIAL ODDSRATIO(LABEL);
     CONTRAST "Multiv Wald test of Model" NLIvsSLI 1, momed12 1;
     RANDOM INTERCEPT / SUBJECT=ID;     * Add per-person "residual" as random intercept;
     COVTEST "Need Extra Variance?" 0;  * Test additive overdispersion;
RUN; TITLE;
```

STATA Output:

```
Mixed-effects GLM                               Number of obs     =         97
Family:                binomial
Link:                     logit
Binomial variable:      ntrials
Group variable:              id                 Number of groups  =         97

                                                Obs per group:
                                                              min =          1
Integration method:     laplace                               avg =        1.0
                                                              max =          1
                                                Wald chi2(2)      =      14.04
Log likelihood = -274.88176                     Prob > chi2       =     0.0009
------------------------------------------------------------------------------
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nlivssli |  -1.793682   .5089051    -3.52   0.000    -2.791118   -.7962467
     momed12 |   .0327918   .1341283     0.24   0.807    -.2300949    .2956784
       _cons |   4.742648   .4350784    10.90   0.000     3.88991    5.595386
-------------+----------------------------------------------------------------
id           |
   var(_cons)|   4.381075   1.028769                      2.765034    6.941619 → extra variance on logit scale
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 981.97      Prob >= chibar2 = 0.0000 → LRT of additive overdispersion
Akaike's information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
           . |         97         .  -274.8818       4    557.7635   568.0624
------------------------------------------------------------------------------
. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =   14.04
         Prob > chi2 =    0.0009
```

SAS Output:

```
           Fit Statistics
-2 Log Likelihood            549.76
AIC  (smaller is better)     557.76
AICC (smaller is better)     558.20
BIC  (smaller is better)     568.10
CAIC (smaller is better)     572.10
HQIC (smaller is better)     561.95

Fit Statistics for Conditional Distribution
-2 log L(Ncorrect | r. effects)      248.93
Pearson Chi-Square                    14.47
Pearson Chi-Square / DF                0.15 → Much lower because extra variance is included in the model

             Covariance Parameter Estimates
                                  Standard
Cov Parm     Subject    Estimate       Error    Gradient
Intercept    ID           4.3810      1.0287    -0.00019 → Extra "residual" variance on logit model-scale
```

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Gradient
Intercept	4.7427	0.4351	Infty	10.90	<.0001	0.000149
NLIvsSLI	-1.7937	0.5089	Infty	-3.52	0.0004	-0.00016
momed12	0.03278	0.1341	Infty	0.24	0.8069	-0.00099

Odds Ratio Estimates

Comparison	Estimate	DF	95% Confidence Limits	
unit change of NLIvsSLI from mean	0.166	Infty	0.061	0.451
unit change of momed12 from mean	1.033	Infty	0.794	1.344

Contrasts

Label	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
Multiv Wald test of Model	2	Infty	14.04	7.02	0.0009	0.0009 → big difference!

Tests of Covariance Parameters Based on the Likelihood

Label	DF	-2 Log Like	ChiSq	Pr > ChiSq	Note
Need Extra Variance?	1	1531.73	981.97	<.0001	MI → LRT with mixture of DF=0,1

---

4) Two-Predictor Model with <u>Multiplicative Over-Dispersion via Beta-Binomial Distribution</u>

$\#Correct_i \sim BetaBinomial(p_i, Ntrials_i, \phi)$ → p_i is still probability of any one trial being correct

$p_i \sim Beta(a_i, b_i)$ → $a_i = p_i/\phi,\ b_i = (1-p_i)/\phi$

$Logit(p_i \text{ for correct trial}) = \beta_0 + \beta_1(NLIvsSLI_i) + \beta_2(MotherEd_i - 12)$

Conditional mean: $\#Correct_i = Ntrials_i * p_i$

Conditional variance of $\#Correct_i = (Ntrials_i * p_i)(1-p_i)[1 + (Ntrials_i - 1)/(\phi + 1)]$

Disclaimer: I struggled to translate this model across different parameterizations I found, and this formula for the conditional variance produced results that were close to those of SAS, but not exactly the same…

```
display as result "STATA Two-Predictor Beta-Binomial Model with Multiplicative Overdispersion"
display as result "Switch to betabin that has beta-binomial distribution"
betabin ncorrect c.nlivssli c.momed12, link(logit) n(ntrials),
estat ic, n(98),
test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
// LRT for multiplicative overdispersion is done for you automatically

display as result "STATA Two-Predictor Beta-Binomial Model with Multiplicative Overdispersion"
display as result "Get Odds Ratios"
betabin ncorrect c.nlivssli c.momed12, link(logit) n(ntrials) eform,
```

```
TITLE1 "SAS Two-Predictor Beta-Binomial Model with Multiplicative Overdispersion";
TITLE2 "Switch to PROC FINITE MIXTURE MODEL that has beta-binomial distribution";
PROC FMM DATA=work.Example3a NAMELEN=100;
     MODEL Ncorrect/Ntrials = NLIvsSLI momed12 / LINK=LOGIT DIST=BETABINOMIAL;
RUN; TITLE1; TITLE2;
```

Note: PROC FMM has far fewer options for post-estimation (no CONTRAST, ESTIMATE, LSMEANS).

STATA Output:

```
Beta-binomial regression                        Number of obs   =         97
Link           = logit                          LR chi2(2)      =      13.61
Dispersion     = beta-binomial                  Prob > chi2     =     0.0035
Log likelihood = -267.05167                     Pseudo R2       =     0.0248
```

```
------------------------------------------------------------------------------
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nlivssli |  -.9737565   .2728606    -3.57   0.000    -1.508553   -.4389595
     momed12 |   .0464046   .0685461     0.68   0.498    -.0879434    .1807525
       _cons |   2.957862   .2500499    11.83   0.000     2.467773     3.44795
-------------+----------------------------------------------------------------
    /lnsigma |  -1.421521   .2207495    -6.44   0.000    -1.854182   -.9888596
-------------+----------------------------------------------------------------
       sigma |   .2413467   .0532772                      .156581    .3720007 = 1/phi multiplier given by SAS
------------------------------------------------------------------------------
Likelihood-ratio test of sigma=0:  chibar2(01) =  997.63 Prob>=chibar2 = 0.000 → LRT for overdispersion

Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+---------------------------------------------------------------
           . |         98  -273.8551  -267.0517       4   542.1033   552.4432
-----------------------------------------------------------------------------
. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =   13.75
         Prob > chi2 =    0.0010
```

SAS Output:

```
           Fit Statistics
-2 Log Likelihood              534.1
AIC  (Smaller is Better)       542.1
AICC (Smaller is Better)       542.5
BIC  (Smaller is Better)       552.4
Pearson Statistic             71.6649 → when divided by DF=96, = 0.75, pretty good!

         Parameter Estimates for Beta-Binomial Model
                                Standard
Effect              Estimate       Error    z Value    Pr > |z|
Intercept             2.9579      0.2500      11.83       <.0001
NLIvsSLI             -0.9738      0.2729      -3.57       0.0004
momed12             0.04640     0.06855       0.68       0.4984
Scale Parameter      4.1434      0.9147                          → phi multiplier for variance (1=binomial?)
```

Btw, I couldn't figure out how to get a multivariate Wald test for the two predictors together (i.e., the model) using PROC FMM ☹

---

5) Two-Predictor Binomial Model with Zero-Inflation (predicting number incorrect now)

Our negatively skewed data have one-inflation, not zero-inflation, but all the software routines I found was designed only for zero-inflation. So I solved this problem by predicting number incorrect instead of number correct. The model below says that number incorrect comes from a binomial distribution that has extra zero values. The "inflation" model that predicts the logit of being an "extra zero" is empty for now, because I just want to see how many there are likely to be.

$$Logit\left(p_{ip} \text{ for incorrect trial}\right) = \beta_{0p} + \beta_{1p}(NLIvsSLI_i) + \beta_{2p}(MotherEd_i - 12)$$

$$Logit(p_{iz} \text{ for } y_i = 0) = \beta_{0z}$$

$$\text{Conditional mean: } \#Incorrect_i = \left(Ntrials_i * p_{ip}\right) * p_{iz}$$

I'm not even going to try to get the distributional notation or conditional variance right…

```
display as result "STATA Two-Predictor Zero-Inflated Binomial Model"
display as result "Switch to zbin and predict Nincorrect"
zib nincorrect c.nlivssli c.momed12, link(logit) n(ntrials) ///
              ilink(logit) inflate(_cons),  // ilink is link for inflate model
estat ic, n(98),
test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model

TITLE1 "SAS Two-Predictor Zero-Inflated Binomial Model";
TITLE2 "Use FMM and predict Nincorrct instead";
PROC FMM DATA=work.Example3a NAMELEN=100;
     MODEL Nincorrect/Ntrials = NLIvsSLI momed12 / LINK=LOGIT DIST=BINOMIAL;
     MODEL + / DIST=CONSTANT; * Inflation model predicting zero;
RUN; TITLE1; TITLE2;
```

STATA Output:

```
Zero-inflated binomial regression               Number of obs   =         97
Regression link: logit                          Nonzero obs     =         57
Inflation link : logit                          Zero obs        =         40
                                                LR chi2(2)      =     126.58
Log likelihood = -494.1091                      Prob > chi2     =     0.0000
------------------------------------------------------------------------------
  nincorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nincorrect   |
    nlivssli |   .6787023   .0934716     7.26   0.000     .4955014    .8619033 Beta1p
     momed12 |  -.1148639    .024894    -4.61   0.000    -.1636552   -.0660727 Beta2p
       _cons |  -2.209937   .0825224   -26.78   0.000    -2.371678   -2.048196 Beta0p
-------------+----------------------------------------------------------------
inflate      |                                                      → logit of being extra 0
       _cons |  -.3547476   .2063317    -1.72   0.086    -.7591502     .049655 Beta0z
------------------------------------------------------------------------------
Akaike's information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
           . |         97 -557.3991  -494.1091       4    996.2183   1006.517
------------------------------------------------------------------------------
               Note: N=97 used in calculating BIC.

. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =  116.04
         Prob > chi2 =    0.0000
```

SAS Output:

```
           Fit Statistics
-2 Log Likelihood              988.2
AIC  (Smaller is Better)       996.2
AICC (Smaller is Better)       996.7
BIC  (Smaller is Better)      1006.5
Pearson Statistic              225.0 → Divided by DF=96, = 2.34375 (not as good)
Effective Parameters               4 → number of parameters here
Effective Components               2 → This is a mixture model

                Parameter Estimates for Binomial Model
                                        Standard
   Component    Effect       Estimate       Error    z Value    Pr > |z|
           1    Intercept     -2.2099     0.08252     -26.78      <.0001 beta0p
           1    NLIvsSLI       0.6787     0.09347       7.26      <.0001 beta1p
           1    momed12       -0.1149     0.02489      -4.61      <.0001 beta2p

              Parameter Estimates for Mixing Probabilities
                           -----------------Linked Scale-----------------
                Mixing                    Standard
Component    Probability    Logit(Prob)       Error    z Value    Pr > |z|

        1         0.5878         0.3547      0.2063       1.72      0.0856
        2         0.4122        -0.3547                                    → Prob and logit of being an extra 0
```

---

6) Two-Predictor <u>Beta-Binomial</u> Model with <u>Zero-Inflation</u> (predicting number incorrect now)

The model below says that number incorrect comes from a beta-binomial distribution that has extra zero values (instead of a binomial distribution that has extra zero values), allowing multiplicative overdispersion.

$Logit(p_{ip} \text{ for incorrect}) = \beta_{0p} + \beta_{1p}(NLIvsSLI_i) + \beta_{2p}(MotherEd_i - 12)$

$Logit(p_{iz} \text{ for } y_i = 0) = \beta_{0z}$

Conditional mean: $\#Incorrect_i = (Ntrials_i * p_{ip}) * p_{iz}$

I'm not even going to try to get the distributional notation or conditional variance right…

```
display as result "STATA Two-Predictor Zero-Inflated Beta-Binomial Model"
display as result "Switch to zibbin and predict Nincorrect"
zibbin nincorrect c.nlivssli c.momed12, link(logit) n(ntrials) ///
                  ilink(logit) inflate(_cons),
estat ic, n(98),
test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
```

```
TITLE1 "SAS Two-Predictor Zero-Inflated Beta-Binomial Model";
TITLE2 "Use FMM and predict Nincorrct instead";
PROC FMM DATA=work.Example3a NAMELEN=100;
     MODEL Nincorrect/Ntrials = NLIvsSLI momed12 / LINK=LOGIT DIST=BETABINOMIAL;
     MODEL + / DIST=CONSTANT; * Inflation model predicting zero;
RUN; TITLE1; TITLE2;
```

STATA Output:

```
Zero-inflated beta-binomial regression           Number of obs   =         97
Regression link: logit                           Nonzero obs     =         57
Inflation link : logit                           Zero obs        =         40
                                                 LR chi2(2)      =      11.61
Log likelihood =  -263.789                       Prob > chi2     =     0.0030
------------------------------------------------------------------------------
  nincorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nincorrect   |
    nlivssli |   1.128224   .3464563     3.26   0.001     .4491825    1.807266 Beta1p
     momed12 |  -.0178967   .0894132    -0.20   0.841    -.1931434    .1573499 Beta2p
       _cons |  -2.750534   .3270209    -8.41   0.000    -3.391483   -2.109585 Beta0p
-------------+----------------------------------------------------------------
inflate      |                                                      → logit of being an extra 0
       _cons |  -1.095397   .4369649    -2.51   0.012    -1.951832   -.2389614 Beta0z
-------------+----------------------------------------------------------------
    /lnsigma |  -1.870879   .2495082                     -2.359906   -1.381852
-------------+----------------------------------------------------------------
       sigma |   .1539883   .0384213                      .0944291    .2511131 → 1/scale multiplier in SAS
------------------------------------------------------------------------------
Akaike's information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
           . |         97 -269.5932   -263.789       5     537.578   550.4516
------------------------------------------------------------------------------
               Note: N=97 used in calculating BIC.

. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =   12.43
         Prob > chi2 =    0.0020
```

Stata would not let me do an LRT to compare the zero-inflated models (even though it should have been possible according to their documentation)… ☹

SAS Output:

```
          Fit Statistics
-2 Log Likelihood              527.6 → -2LL diff = 460.0 relative to zero-inflated binomial, so is better
AIC  (Smaller is Better)       537.6
AICC (Smaller is Better)       538.2
BIC  (Smaller is Better)       550.5
Pearson Statistic             84.2707 → Divided by DF=96, = 0.878 (better)
Effective Parameters                5 → number of parameters here
Effective Components                2 → still a mixture model

                Parameter Estimates for Beta-Binomial Model
                                              Standard
   Component    Effect            Estimate       Error    z Value    Pr > |z|
           1    Intercept          -2.7505      0.3270      -8.41      <.0001 beta0p
           1    NLIvsSLI            1.1282      0.3465       3.26      0.0011 beta1p
           1    momed12           -0.01789     0.08941      -0.20      0.8414 beta2p
           1    Scale Parameter     6.4940      1.6203                         → phi multiplier is bigger now
```

```
              Parameter Estimates for Mixing Probabilities
                         -----------------Linked Scale-----------------
                Mixing                    Standard
Component    Probability    Logit(Prob)       Error    z Value    Pr > |z|
        1         0.7494         1.0954      0.4370       2.51       0.0122
        2         0.2506        -1.0954                                    → Prob and Logit of being an extra 0
```

---

7) Four-Predictor Beta-Binomial Model with Zero-Inflation (now predictors in inflation model)

The model below adds our two predictors to the zero-inflation model (customizing probability of being an extra zero).

$Logit(p_i \text{ for incorrect}) = \beta_{0p} + \beta_{1p}(NLIvsSLI_i) + \beta_{2p}(MotherEd_i - 12)$
$Logit(p_{iz} \text{ for } y_i > 0) = \beta_{0z} + \beta_{1z}(NLIvsSLI_i) + \beta_{2z}(MotherEd_i - 12)$
Conditional mean: $\#Incorrect_i = (Ntrials_i * p_i) * p_{iz}$

I'm not even going to try to get the distributional notation or conditional variance right…

```
display as result "STATA Two-Predictor Zero-Inflated Beta-Binomial Model"
display as result "Switch to zibbin and predict Nincorrect"
zibbin nincorrect c.nlivssli c.momed12, link(logit) n(ntrials) ///
                  ilink(logit) inflate c.nlivssli c.momed12),
estat ic, n(98),
test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
```

STATA Output only (SAS PROC FMM wouldn't allow zero-model predictors):

```
Zero-inflated beta-binomial regression           Number of obs   =         97
Regression link: logit                           Nonzero obs     =         57
Inflation link : logit                           Zero obs        =         40
                                                 LR chi2(2)      =       7.38
Log likelihood = -261.8274                       Prob > chi2     =     0.0249
------------------------------------------------------------------------------
  nincorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nincorrect   |
    nlivssli |   .3036772   .3546852     0.86   0.392     -.391493    .9988474 Beta1p
     momed12 |  -.2189386   .0812336    -2.70   0.007    -.3781535   -.0597237 Beta2p
       _cons |  -2.173967   .3963158    -5.49   0.000    -2.950731   -1.397202 Beta0p
-------------+----------------------------------------------------------------
inflate      |                                                                 → predict logit of extra 0
    nlivssli |  -3.970179   5.512301    -0.72   0.471    -14.77409    6.833733 Beta2z
     momed12 |  -.9569979   1.428802    -0.67   0.503    -3.757398    1.843402 Beta1z
       _cons |   .0198758   .6209887     0.03   0.974     -1.19724    1.236991 Beta0z
-------------+----------------------------------------------------------------
    /lnsigma |  -1.652934   .3139631                      -2.26829   -1.037578
-------------+----------------------------------------------------------------
       sigma |   .1914873   .0601199                      .103489     .354312 → 1/scale in SAS
------------------------------------------------------------------------------

Akaike's information criterion and Bayesian information criterion
------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
           . |         97 -265.5186  -261.8274       7    537.6548   555.6777
------------------------------------------------------------------------------
```

So which one should be pick? Let's do some informal model comparisons using distribution fit and relative fit (*may not be exactly comparable due to differences in estimation technique, but they should be close)

Two-Predictor Model	Pearson Chi-Square / DF	AIC*	BIC*
2. Regular Binomial	15.41	1537.7	1545.5
3. +Additive Overdispersion	0.15	557.8	568.1
4. Beta-Binomial	0.75	542.1	552.4
5. Zero-Inflated Binomial	2.34	996.2	1006.5
6. Zero-Inflated Beta-Binomial	**0.88**	**537.6**	**550.5**
7. ZIBB + Predictors	?	537.7	555.7

Sample Write-up using both programs (final model = zero-inflated beta-binomial without inflation predictors):

The extent that grammatical understanding (measured either as percent correct or percent incorrect; see below) at third grade could be predicted by language impairment group (non-specific=0, specific=1) and mother's years of education (centered such that 0=12 years) was examined in a series of generalized linear models. In the sample of $N = 97$ children, the mean percent correct was 0.92, with a large percentage of observations at or near the ceiling (1.00). Accordingly, we predicted the number of correct trials out of the number of possible trials using a logit link function to keep the predicted percent correct outcomes below 1. The type of model specifies that the number of correct responses follows a binomial-based distribution with 100 total trials and a model-predicted probability of a correct response on any trial. While the model predicts the logit (log-odds) of a correct answer for any trial, that prediction can be translated into percent correct via an inverse link function (which provides model-predicted proportions and their standard errors). All models were estimated using maximum likelihood within SAS GLIMMIX and FMM to assess distribution fit, as well as in stata glm, betabin, zib, and zibbin; predictor fixed effects were tested univariately using z-distributions without denominator degrees of freedom. Effect sizes are provided below as odds ratios: the exponentiated logit coefficient in which values from 0 to 1 indicate negative associations, 1 indicates no association, and values above 1 indicate positive associations.

Before interpreting our results, we tested the fit of models with alternative binomial-based conditional outcome distributions (each with main effects of group and mother's education) by examining the Pearson χ^2/DF statistic (in which 1=good fit), as well as likelihood ratio tests (i.e., treating −2 times the difference in log-likelihood between nested models as a χ^2 statistic with degrees of freedom equal to the number of additional parameters). As expected given the negatively skewed observed distribution, a model specifying a standard binomial distribution for number correct did not fit well (Pearson $\chi^2/DF = 15.41$). Two methods of allowing overdispersion were then examined. First, we allowed additive overdispersion via an observation-level random intercept, which significantly improved model fit, $-2\Delta LL(1) = 987.97$, $p < .0001$, but created a tendency towards underdispersion (Pearson $\chi^2/DF = 0.15$). Second, we allowed multiplicative overdispersion by using a beta-binomial distribution, which significantly improved model fit, $-2\Delta LL(1) = 997.63$, $p < .0001$, and appeared to fit well (Pearson $\chi^2/DF = 0.75$). We then examined the potential for one-inflation by predicting number *incorrect* instead so that zero-inflation models could be fitted. A model predicting number incorrect with a zero-inflated binomial distribution was examined but did not fit as well (Pearson $\chi^2/DF = 2.34$), although using a zero-inflated beta-binomial distribution instead did result in good fit (Pearson $\chi^2/DF = 0.88$), as well as the lowest AIC and BIC of all the models. We also examined group and mother's education as predictors of zero-inflation but neither was significant (with higher AIC and BIC values), and thus the empty (unconditional) zero-inflation model was retained.

The model results indicated that 25.06% of the sample were predicted to be an extra 0 (i.e., to be part of the zero-inflated part of the distribution for number incorrect). Otherwise, the predicted intercept for a child with non-specific language impairment whose mother had 12 years of education was a logit = −2.75, which translates into percent incorrect = 0.06. Children with specific language impairment were predicted to have significantly more incorrect responses (logit = 1.12, OR = 3.09), although no significant difference was found for mother's years of education (logit = −0.02, OR = 0.98). The scale parameter for multiplicative overdispersion was 6.494, which was significant, $-2\Delta LL(1) = 460.60$, $p < .0001$.