Does My Baby Really Look Like Me? Using Tests For Resemblance Between Parent and Child to Teach Topics in Categorical Data Analysis

Amy G. Froelich

Dan Nettleton

Iowa State University

Abstract

In this article, we present a study to test whether neutral observers perceive a resemblance between a parent and a child. We demonstrate the general approach for two separate parent/child pairs using survey data collected from introductory statistics students serving as neutral observers. We then present ideas for incorporating the study design process, data collection, and analysis into different statistics courses from introductory to graduate level.

1. Introduction

Many new parents have heard claims of a striking resemblance between them and their babies. As new parents ourselves, we were skeptical of such claims so we devised a study to objectively evaluate resemblance in a particular parent/child pair. In this article, we discuss the methods used in our study, the format of the data collection instruments, and how the

data from this study can be collected in a classroom setting and used to illustrate several statistical concepts in categorical data analysis ranging in complexity from introductory to graduate level.

While our focus is on assessing evidence of resemblance for a particular parent/child pair, other researchers have addressed the more general question of whether children tend to resemble their fathers, their mothers, or both. Christenfeld and Hill (1995), for example, collected data on parent/child resemblance using 24 families, each consisting of a father, a mother, and their child. Neutral judges were presented with a picture of each of the 24 children and asked to guess which of three pictured fathers and which of the three pictured mothers was the actual biological parent. Christenfeld and Hill found that, in most cases, judges were not able to match children to their fathers or mothers at a rate significantly greater than would be expected under random guessing. The one exception was that one-year-old children were matched to their fathers at a rate significantly higher than expected by chance. Christenfeld and Hill suggested a possible evolutionary rationale: a resemblance between a baby and a father may enhance paternal investment in child care by assuring the father that the baby is his. A resemblance between the baby and the mother would not be similarly advantageous because the mother (having given birth to the baby) can be sure the baby is hers.

Several other studies have reached conclusions that differ from Christenfeld and Hill's in various ways. Alvergne, Faurie, and Raymond (2007) described several flaws in previous study designs that could explain these discrepancies. Several of these studies cited by Alvergne et al. used low quality pictures from family albums, or used different types of pictures (casual vs. formal), making it difficult for judges to detect resemblance. In some studies, the pictures used had similar backgrounds for both the child and the parent(s), giving judges visual clues other than resemblance for some of the parent/child pairs. Alvergne et al. also noted that several studies used a fixed set of foils (incorrect parents) for each true parent/child pair instead of randomly selecting foils from a larger population. The use of a fixed set of foils for each parent/child pair tests the resemblance of only the pictured sets of fathers or mothers to the child, instead of testing for general parent/child resemblance. Based on their own careful study, Alvergne et al. conclude that children from birth to age six tend to resemble both their fathers and mothers more than would be expected by chance and that the parent to which the resemblance is stronger depends on the age and sex of the child.

We found many studies in the literature focused on the general question of resemblance, for example, between parents and children (Alvergne et al. 2007; Christenfeld and Hill 1995), dogs and owners (Roy and Christenfeld 2004; Levine 2005; Roy and Christenfeld 2005), and husbands and wives and wives and mothers-in-law (Bereczkei et al. 2002). We

were unable to find a study addressing only resemblance between a particular parent and his or her child. Because this is the focus of our research, our study design and research questions (Section 2), data collection surveys (Section 3), and analysis strategies (Section 4) differ from those in the literature. After presenting our study, we discuss ideas for how instructors can incorporate aspects of this study into specific statistics courses in Section 5. Conclusions follow in Section 6.

2. Study Design and Research Questions

The goal for our study is to test for the resemblance specifically between two parent/child pairs consisting of the first author and her daughter and the second author and his son. We also want to collect data for this study using students from our introductory statistics courses as judges and then use the data to motivate topics in the introductory and higher level statistics courses. At the same time, we wish to avoid, if possible, some of the problems noted by Alvergne et al. (2007) in previous studies on general parent/child resemblance.

Since the focus of our study is on determining if our children look like us, for each parent/child pair, we presented a picture of the parent and asked judges (students) to guess which of four babies pictured is the actual baby of the parent. As in other studies of this nature, selection of any of the babies at a rate higher than expected by chance (1/4) indicates the judges detect a resemblance between the parent and that baby.

Thus, for each parent/child pair, our method requires five pictures, one for the parent and one for each of the four babies. The parent picture was taken in front of a plain background with no additional visual clues present. The four babies for each parent/child pair were pictured at approximately the same age (six months) and in similar settings (solo studio pictures). These choices allow us to control for the quality of the pictures of the babies and for the backgrounds in the pictures of the babies and the parent. Care was taken to select three other babies with similar physical characteristics to the child (e.g., same race and gender) but not directly related to the parent (other children or nieces/nephews of the parent were not used.) In order to facilitate comparison between the four babies, the pictures were presented together in a 2×2 pattern and labeled A through D. Placement of the pictures in this pattern was randomly determined for each set of four babies and then held constant throughout.

Random selection of the three other babies (the foils) from a larger population of babies is not practical for a study of this nature. Possible foils were selected from our own small collection of photographs. Once pictures of children with dissimilar characteristics were

eliminated, we were left with only a few more pictures than needed for the study. Randomizing the foils would require a much larger number of available baby pictures and would greatly complicate or eliminate many of the data collection methods available in the classroom, such as surveys through the university's course management system or questions in class with personal response systems (clickers). Thus, since each judge (student) evaluated resemblance from the same set of four baby pictures, all conclusions here about the resemblance of each parent/child pair are made relative to the three other babies pictured.

Other studies on resemblance (Alvergne, Faurie, and Raymond 2007; Christenfeld and Hill 1995) also include demographic variables on the judges in their analyses. For example, Alvergne et al. included the variables gender, age, and both the number of children and the number of siblings of the judge. After the use of typical variable selection methods, none of these variables were included in the final model in their analysis. Given the use of college students as the judges in our study, variables like age and number of children tend to vary little and would likely be uninformative. Thus, such demographic variables were not considered in our study. We, however, did record the self-reported gender of each judge so that any impact of this factor on the response could be evaluated.

Based on the design of our study, we developed four sets of research questions concerning the resemblance of the two parent/child pairs. The first set of research questions concerns whether or not the judges detect a resemblance between the parent and one of the babies pictured.

- **Q1a:** For each parent/child pair, when presented with a picture of the parent and a set of four baby pictures, do judges detect a resemblance between the parent and any of the babies pictured?

- **Q1b:** For each parent/child pair, when presented with a picture of the parent and a set of four baby pictures, is the gender of the judge associated with the baby selected?

The second set of research questions concerns whether or not the judges detect a resemblance between the parent and his/her child.

- **Q2a:** For each parent/child pair, when presented with a picture of the parent and a set of four baby pictures, do judges detect a resemblance between the parent and his/her baby?

- **Q2b:** For each parent/child pair, when presented with a picture of the parent and a set of four baby pictures, does the probability of selecting the correct baby depend on the gender of the judge?

The third research question concerns the degree of resemblance detected between the parent/child pair. In order to claim the parent looked more like his/her baby than each of the other three babies pictured, the judges not only need to select the correct baby at a rate higher than expected based on random selection, they need to select the correct baby significantly more often than each of the other babies.

- **Q3:** For each parent/child pair, when presented with a picture of the parent and a set of four baby pictures, do the judges select the correct baby more frequently than each of the other babies pictured?

The fourth and final set of research questions includes questions specific to our two parent/child pairs. For the parent/child pair of the first author and her daughter, several family members had commented on the resemblance between her daughter and the first author's baby pictures. To determine if the resemblance between the first author and her daughter would be easier to see when judges view pictures of the two taken at the same age, we compare the responses of judges when viewing a picture of the first author as an adult versus viewing a picture of her as a baby. For the parent/child pair of the second author and his son, the son was the only baby pictured wearing a hat. Instead of selecting his son based on resemblance, the judges could select his son based on this difference. Thus, for this parent/child pair, we want to investigate the degree to which baby choice might be influenced by factors other than resemblance with the father.

- **Q4a:** Do judges make consistent baby selections when viewing a picture of the first author as an adult, versus when viewing a picture of the first author as a baby? Which selection, if either, is more accurate?

- **Q4b:** Are judges influenced by a factor present in the baby pictures (e.g., baby wearing a hat) other than resemblance to the parent?

3. Surveys and Data Collection

To determine answers to the four sets of research questions, we developed four surveys, two for each parent/child pair. The first survey for each parent/child pair contains two questions and is designed to answer research questions 1 through 3. The first question asks the gender of the judge and the second question asks the judge to guess which of the four babies pictured is the child of the parent whose picture is also displayed on the screen. These surveys are referenced as Survey MD1 for the parent/child pair of the first author and her daughter and Survey FS1 for the parent/child pair of the second author and his son.

The second survey for each parent/child pair (called Survey MD2 and Survey FS2) contains three questions and is designed to answer research question 4. For both surveys, the first question asks the gender of the judge. In Survey MD2, the other two questions ask the judge to select the baby of the parent pictured first as an adult (in the second question) and then as a baby (in the third question). This allows us to examine the consistency of responses between the two pictures of the first author (adult vs. baby) and determine which selection, if either, is more accurate. In Survey FS2, the second question shows only the four baby pictures and asks the judges to guess the baby *before* seeing the father. Thus, the judges' selections are based entirely on factors other than resemblance between father and son. In the third question, the parent is pictured, allowing us to look at both parent/child resemblance and consistency of responses between the two selections.

Since one of the secondary goals of this study is to collect data to motivate certain topics in introductory statistics, we invited students enrolled in three different introductory courses at a large Midwestern university during a recent semester to participate in this study. Surveys were administered electronically through the university's course management system, and all survey responses were recorded anonymously. Each student was randomly assigned to take one survey for each parent/child pair based on the last digit of their university identification number. Due to the structure of Surveys MD2 and FS2, questions were administered one at a time, and students were not allowed to revisit previously answered questions. As a part of the university's Institutional Review Board approval of this project, students did not receive compensation for completing the survey, and course instructors were not informed of student participation in the project. Surveys were open for one week, after which time, students were informed of the correct parent/child pairs.

Two hundred twenty students completed Survey MD1, and 140 students completed Survey FS1. While all students completing these two surveys answered the second question on baby selection, a small number (7 for Survey MD1 and 8 for Survey FS1) did not indicate their gender. The responses from these 15 students are excluded only from analyses which include gender. One hundred twenty six students completed Survey MD2, and 203 students completed Survey FS2. However, only responses from the 123 and 192 students who answered both the second and third questions on Survey MD2 and FS2, respectively, are included in the analysis.

4. Data Analysis

In this section, we use the data collected from the four surveys described in Section 3 to answer the four sets of research questions from Section 2. The methods in Sections 4.1 and 4.2 are typically covered in introductory statistics courses and can be found in textbooks

such as Moore, McCabe, and Craig (2009) or DeVeaux, Velleman, and Bock (2009). The methods in Sections 4.3 through 4.4 are more appropriate for an upper level undergraduate or graduate level course in categorical data analysis. The methodology in Section 4.3 can be found in Nettleton (2009). The methods described in Section 4.4 can be found in textbooks such as Agresti (2007).

4.1 Research Question 1

Our first set of research questions concerns whether or not judges detect a resemblance between the parent and at least one of the babies pictured and if the pattern of baby selection depends on the gender of the judge. For research question Q1a, if the judges do not detect a resemblance between the parent and any of the babies, the probability of each baby being selected is 0.25. However, if the judges do detect a resemblance, the probability for at least one baby will be different than 0.25. For $j = A, B, C, D$; let n_j denote the number of judges who select baby j. Under the null hypothesis of equal probability of selection, the expected count for each baby is $0.25n$, where n is the total number of judges. The relevant goodness of fit test statistic is therefore

$$X^2 = \sum_{j=A}^{D} \frac{(n_j - 0.25n)^2}{0.25n}. \tag{1}$$

For our large samples, the X^2 statistic will have an approximate χ^2 distribution with 3 degrees of freedom under the null.

Figure 1 shows the number of judges who selected baby A through D in Survey MD1, where baby C is the correct choice, and in Survey FS1, where baby B is the correct choice. For Survey MD1, these counts differ statistically from those expected under the null hypothesis of equal multinomial probabilities ($X^2 \approx 74.4132$, p-value ≈ 0). Based on the data, the judges see a resemblance between the first author and either baby B or baby C. For Survey FS1, these counts also differ statistically from those expected under the null ($X^2 \approx 21.5429$, p-value ≈ 0.00008). Clearly, in this case, the judges perceive a resemblance between the second author and baby D.

For both parent/child pairs, evidence of resemblance between the parent and at least one of the babies is present. If the gender of the judge is not associated with the baby selected, the proportion of judges selecting each baby should be approximately the same for each gender. If there is an association, the probabilities of baby selection will vary between the two genders. Therefore, to answer research question Q1b, we can test for the equality of two multinomial probability vectors. Let $n_{g,j}$ denote the number of judges in the g, j^{th} cell of the contingency table, where g denotes either Females (F) or Males (M) and j is the baby selected (A through D). As before, denote the overall number of judges who select baby j
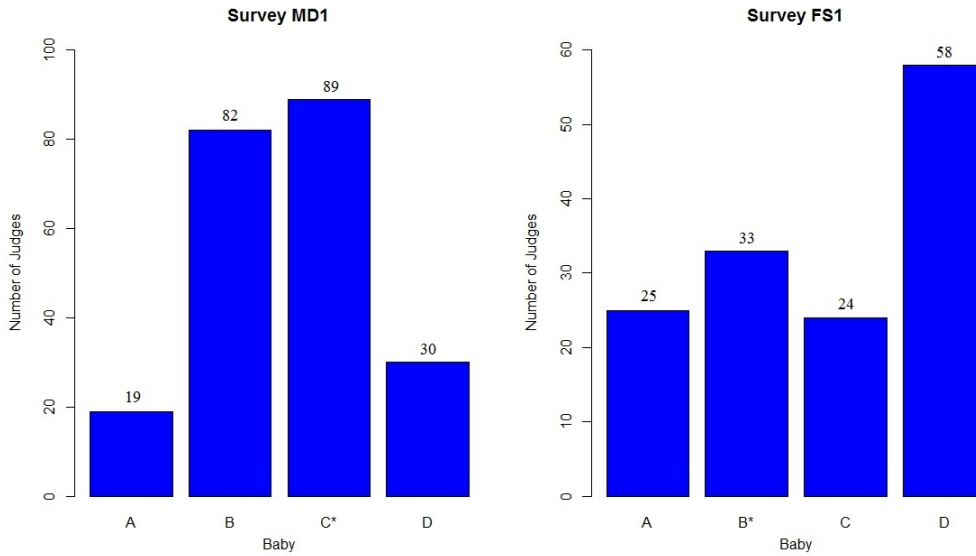
Figure 1.   Number of judges selecting Babies A through D in surveys MD1 and FS1.

as n_j and denote the number of female and male judges as n_F and n_M respectively. Under the null hypothesis, the baby selection probabilities are the same for males and females, and the expected number of judges in the g, j^{th} cell of the contingency table is $n_g n_j / n$. The Pearson χ^2 test statistic is then

$$X^2 = \sum_g \sum_j \frac{(n_{g,j} - n_g n_j / n)^2}{n_g n_j / n}. \tag{2}$$

For our large samples, the test statistic X^2 has an approximate χ^2 distribution with 3 degrees of freedom under the null hypothesis.

The contingency tables of baby selected and gender of judge are given in Table 1 for Survey MD1 and in Table 2 for Survey FS1. For Survey MD1, the test statistic from Equation (2) is $X^2 \approx 0.482$ with p-value ≈ 0.9229. For Survey FS1, the test statistic from Equation (2) is $X^2 \approx 2.970$ with p-value ≈ 0.3963. Thus, we find no evidence of a gender difference in baby selection probabilities for either parent/child pair.

Table 1.   Number of Judges Selecting Babies *A* through *D* by Gender in Survey MD1

Gender	Baby Selected				Total
	A	*B*	*C*	*D*	
Female	10	42	45	14	111
Male	7	39	41	15	102
Total	17	81	86	29	213

Table 2.   Number of Judges Selecting Babies *A* through *D* by Gender in Survey FS1

Gender	Baby Selected				Total
	A	*B*	*C*	*D*	
Female	14	17	15	26	72
Male	9	13	8	30	60
Total	23	30	23	56	132

4.2   Research Question 2

The second set of research questions concerns whether or not the judges detect a resemblance between the parent/child pair, and whether the probability of a correct response depends on the gender of the judge. If a judge is not able to detect the resemblance in a parent/child pair and is simply randomly selecting each child with equal probability, the probability of a correct response would be 0.25. On the other hand, if there is a special resemblance between parent and child detected by the judges, the probability of a correct response would be greater than 0.25. If we let X_i equal 1 if the ith judge selects the correct baby and 0 otherwise for $i = 1, \ldots, n$ judges, it is natural to model $X_1, \ldots, X_n$ as independent and identically distributed Bernoulli random variables with success probability p. To answer research question Q2a, we wish to test

$$H_0 : p = 0.25 \text{ vs. } H_A : p > 0.25.$$

Though there are a variety of approaches for conducting this one-sided test, introductory statistics courses use a one-sample z-test for a binomial proportion, which is a special case of a score test. The test statistic, based on the sample proportion of correct responses ($\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$), is

$$z = \frac{\hat{p} - 0.25}{\sqrt{\frac{0.25(0.75)}{n}}}. \tag{3}$$

The test statistic z is approximately standard normal under H_0 for our sample sizes.

From Figure 1 for Survey MD1, $\hat{p} = 89/220 \approx 0.4045$ and $z \approx 5.2938$, yielding a p-value of approximately 0. Thus, there is strong evidence the judges detect a resemblance between the first author and her daughter. However, from Figure 1 for Survey FS1, $\hat{p} = 33/140 \approx 0.2357$. Because $\hat{p} < 0.25$, there is no evidence the judges detect a resemblance between the second author and his son.

For research question Q2b, we wish to test whether or not the probability of a correct response is different for each gender. Did both genders perform equally well (as in Survey

MD1) or equally poorly (as in Survey FS1) in selecting the correct baby? We can answer this question using a two-sample test for the equality of binomial proportions. Let the sample proportion of correct responses over all judges be $\hat{p}_{\text{pooled}}$ and let the sample proportion of correct responses for female and male judges be $\hat{p}_F$ and $\hat{p}_M$, respectively. Again considering a score test, the z-test statistic for the equality of the two binomial proportions is

$$z = \frac{\hat{p}_F - \hat{p}_M}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_F} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_M}}}, \quad (4)$$

where n_F and n_M are the number of female and male judges, respectively. Under the null hypothesis of equality of gender-specific success probabilities, the test statistic in Equation (4) will have approximately a standard normal distribution for our sample sizes and proportions.

From Table 1 for Survey MD1, the sample proportion $\hat{p}_F = 45/111 \approx 0.4054$ and the sample proportion $\hat{p}_M = 41/102 \approx 0.4020$. The z-test statistic from Equation 4 is approximately 0.0512 with a corresponding p-value of 0.9592. Thus, we find no evidence of a gender difference in ability to detect the resemblance between the first author and her daughter. A similar analysis of the data from Table 2 ($\hat{p}_F = 17/72$, $\hat{p}_M = 13/60$, $z \approx 0.2654$, p-value ≈ 0.7907) yields the analogous conclusion for Survey FS1.

4.3 Research Question 3

Based on the answer to research question Q2a, judges did not detect a resemblance between the second author and his son. However, for the parent/child pair of the first author and her daughter, judges chose the correct baby in Survey MD1 at a rate (89/220 from Figure 1) significantly greater than that expected under random guessing (1/4). This provides strong evidence the judges perceive a resemblance between mother and daughter. However, another baby was selected by 82 of the 220 judges. This selection rate is also significantly greater than 1/4, which raises a natural question: is the probability of selecting the correct baby greater than the selection probability for each of the other babies? If not, we cannot claim that the mother looks more like her daughter than each of the other three babies.

To answer research question Q3, let n_j again denote the number of judges who selected baby j for $j = A, B, C, D$. We assume that $(n_A, n_B, n_C, n_D)'$ is distributed as a multinomial random vector with $n = \sum_{j=A}^{D} n_j$ trials and cell probabilities p_A, p_B, p_C, and p_D. Let $\hat{p}_j = n_j/n$ for $j = A, B, C, D$. Given that baby C is the daughter of the first author, we wish to test

$$H_0 : p_C \leq p_j \text{ for some } j = A, B, D \text{ vs. } H_A : p_C > p_j \text{ for all } j = A, B, D. \quad (5)$$

Nettleton (2009) derived the likelihood ratio test of Equation (5), showed that the likelihood

ratio test is equivalent to an Intersection Union Test (IUT), and developed several other IUTs for this testing problem. In this case, the simplest of the procedures proposed by Nettleton (2009) rejects H_0 at level $\alpha \in (0,1)$ if and only if separate Wald tests of $p_C \leq p_A$, $p_C \leq p_B$, and $p_C \leq p_D$ are each rejected at level α. For $j = A, B, D$; the Wald statistic for testing $p_C \leq p_j$ is given by

$$
\begin{aligned}
W_j &= \frac{\hat{p}_C - \hat{p}_j}{\sqrt{\widehat{\text{var}}(\hat{p}_C - \hat{p}_j)}} \\
&= \frac{\hat{p}_C - \hat{p}_j}{\sqrt{\hat{p}_C(1 - \hat{p}_C)/n + \hat{p}_j(1 - \hat{p}_j)/n + 2\hat{p}_C\hat{p}_j/n}} \\
&= \frac{\sqrt{n}(\hat{p}_C - \hat{p}_j)}{\sqrt{\hat{p}_C + \hat{p}_j - (\hat{p}_C - \hat{p}_j)^2}}.
\end{aligned}
$$

Each Wald statistic W_j is asymptotically standard normal when $p_C = p_j$. Thus, an approximate one-sided p-value is given by $1 - \Phi(W_j)$ for each $j = A, B, D$; where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The multinomial response vector for Survey MD1 from Figure 1 is $(19, 82, 89, 30)'$. Thus, $W_B \approx 0.536$, and a one-sided p-value for the test of $p_C \leq p_B$ is approximately 0.296. It follows from Nettleton (2009) that we cannot reject the null hypothesis in (5) at significance levels below 0.296, and thus we cannot conclude that the first author looks more like her daughter than each of the other babies pictured.

4.4 Research Question 4

In addition to detecting the resemblance of a parent/child pair, we are interested in looking at two additional aspects of baby selection and resemblance specific to our two parent/child pairs. In Survey MD2, we want to determine if neutral observers detect resemblance at a different rate when the parent is pictured as an adult versus pictured as a baby. In Survey FS2, we want to determine if factors other than resemblance influence the baby selected by the judges.

Therefore, both Survey MD2 and FS2 include two questions asking judges to select one of the four babies pictured. Let X_{i1} be 1 if the ith judge selects the correct baby on the first baby selection question and 0 otherwise. Define X_{i2} in the same way for the second baby selection question. We model $X_{1k}, \ldots, X_{nk}$ as independent and identically distributed Bernoulli random variables with success probability p_k for $k = 1, 2$. We assume that X_{i1} and X_{j2} are independent for all $i \neq j$ but allow X_{i1} and X_{i2} to be dependent because of the likely dependence between the two guesses from each judge. To answer research questions Q4a and Q4b, we are interested in testing for the equality of the binomial probabilities p_1 and

p_2. Because the same n judges provided the responses to both questions, the two sample proportions of correct responses $\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{i1}$ and $\hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{i2}$ are correlated. Thus, the method for testing for the equality of two binomial proportions discussed in Section 4.2 is not appropriate.

We can instead test for the equality of these binomial proportions by using McNemar's test (McNemar 1947). Let n_{CI} denote the number of judges with the response pattern ($X_{i1} = 1, X_{i2} = 0$) and let n_{IC} denote the number of judges with the response pattern ($X_{i1} = 0, X_{i2} = 1$). Thus, n_{CI} and n_{IC} are the entries in the off-diagonal cells of the contingency table of correct and incorrect responses to the two baby selection questions. McNemar's test statistic for the equality of the binomial proportions p_1 and p_2 is

$$z_0^2 = \frac{(n_{CI} - n_{IC})^2}{(n_{CI} + n_{IC})}. \tag{6}$$

Under the null hypothesis of equal binomial probabilities ($p_1 = p_2$), the test statistic z_0^2 will have an approximate χ^2 distribution with 1 degree of freedom for our sample sizes.

For Survey MD2, the contingency table of correct and incorrect responses to the two baby selection questions is given in Table 3. McNemar's test statistic from Equation 6 is $z_0^2 \approx 7.0435$ with corresponding p-value ≈ 0.0080. Thus, we conclude the proportion of correct responses for the two baby selection questions did differ across questions, with judges selecting the correct baby more often when shown the adult picture of the parent than when subsequently shown the baby picture of the parent. With the parent pictured as an adult, the sample proportion of correct responses ($\hat{p}_1 = 54/123 \approx 0.4390$) is larger than expected from random guessing ($z \approx 4.841$, p-value ≈ 0 using the method from Section 4.2), which is consistent with the result from Survey MD1. However, with the parent pictured as a baby, the sample proportion of correct responses ($\hat{p}_2 = 36/123 \approx 0.2927$) is not significantly larger than the random guessing probability of 0.25 ($z \approx 1.0932$, p-value ≈ 0.1371).

Table 3. Contingency Table of Correct and Incorrect Responses for the Two Baby Selection Questions from Survey MD2

With Parent Pictured as Adult	With Parent Pictured as Baby		Total
	Correct	Incorrect	
Correct	22	32	54
Incorrect	14	55	69
Total	36	87	123

For Survey FS2, the contingency table of correct and incorrect responses to the two baby selection questions is given in Table 4. McNemar's test statistic from Equation 6 is $z_0^2 \approx 14.2222$ with corresponding p-value ≈ 0.0002. We conclude that the proportion of correct

responses for the two baby selection questions differs across questions, with the judges selecting the correct baby more often when *not* shown the picture of the parent than when shown the picture of the parent. Judges may be influenced in their baby choice by an outside factor (baby wearing a hat) since the sample proportion of correct responses for the question without the parent pictured ($\hat{p}_1 = 66/195 \approx 0.3385$) is larger than expected under random guessing ($z \approx 2.8528$, *p*-value ≈ 0.0022 using the method from Section 4.2). However, in this case, this influence does not extend to actually detecting father/son resemblance since the sample proportion of correct responses for the question with the parent pictured is $\hat{p}_2 = 34/195 \approx 0.1744 < 0.25$ as in Survey FS1.

Table 4. Contingency Table of Correct and Incorrect Responses for Two Baby Selection Questions from Survey FS2

	With Parent Pictured		
Without Parent Pictured	Correct	Incorrect	Total
Correct	14	52	66
Incorrect	20	109	129
Total	34	161	195

We can also answer research questions Q4a and Q4b by testing for the equality of the multinomial probability vectors for each baby selection question. Let n_{jk} be the number of judges selecting baby $j \in \{A, \ldots, D\}$ on the first baby selection question and baby $k \in \{A, \ldots, D\}$ on the second baby selection question. For $j = A, \ldots, D$; let $n_{j\cdot}$ denote the number of judges selecting baby j on the first question, and let $n_{\cdot j}$ denote the number of judges selecting baby j on the second question. We assume that $(n_{A\cdot}, \ldots, n_{D\cdot})'$ follows a multinomial distribution with cell probabilities $p_{A\cdot}, \ldots, p_{D\cdot}$. Likewise, we assume that $(n_{\cdot A}, \ldots, n_{\cdot D})'$ has a multinomial distribution with cell probabilities $p_{\cdot A}, \ldots, p_{\cdot D}$. We are interested in testing

$$H_0 : p_{j\cdot} = p_{\cdot j} \text{ for all } j = A, \ldots, D \text{ vs. } H_A : p_{j\cdot} \neq p_{\cdot j} \text{ for some } j = A, \ldots, D. \quad (7)$$

The same *n* judges provided the responses for both questions, making the method for testing for the equality of two multinomial probability vectors in Section 4.1 inappropriate. Instead, we can test for the equality of the two multinomial probability vectors by using an extension of McNemar's test (Stuart 1955). Let the vector $\hat{d}$ be the difference in the observed multinomial proportions $\hat{d}_j = (n_{j\cdot} - n_{\cdot j})/n$ for $j = A, B, C$. The difference in the last category D is not included since $\hat{d}_D = -\sum_{j=A}^{C} \hat{d}_j$. Under the null hypothesis (7), $E(\hat{d}) = 0$. The score test statistic is

$$W_0 = n\hat{d}' \hat{V}_0^{-1} \hat{d}, \quad (8)$$

13

where the matrix $\hat{V}_0$ has elements

$$\hat{v}_{jj} = (n_{j\cdot} + n_{\cdot j} - 2n_{jj})/n$$
$$\hat{v}_{jk} = -(n_{jk} + n_{kj})/n \qquad \text{for } j \neq k.$$

Under the null hypothesis in (7), W_0 has an approximate χ^2 distribution with 3 degrees of freedom for our sample sizes.

For Survey MD2, the contingency table of the baby selected for each of the baby selection questions is given in Table 5. For both questions, baby C is the correct choice. The value of the test statistic from Equation (8) is $W_0 \approx 29.1026$ with corresponding p-value ≈ 0. This indicates the marginal baby selection probabilities from the two questions differ for Survey MD2. A study of the marginal distributions shows the judges initially saw a resemblance between the adult picture of the mother and babies B and C (similar to the results from Survey MD1), but once the mother was pictured as a baby, judges tended to switch to baby A (an incorrect choice).

Table 5.   Contingency Table of Baby Selection for Two Baby Selection Questions on Survey MD2

Baby Selected With Adult Parent Picture	Baby Selected With Baby Parent Picture				
	A	B	C	D	Total
A	7	1	4	0	12
B	18	12	8	5	43
C	17	7	22	8	54
D	3	2	2	7	14
Total	45	22	36	20	123

For Survey FS2, the contingency table of the baby selected for each of the baby selection questions is given in Table 6. For both questions, baby B is the correct choice. The value of the test statistic from Equation (8) is $W_0 \approx 24.2329$ with corresponding p-value ≈ 0.00002. Thus, there is evidence the marginal baby selection probabilities differ for the two questions. The multinomial probabilities for the four babies are not all equal to $1/4$ when the parent is not pictured ($X^2 \approx 14.1692$, p-value ≈ 0.0027 using the method discussed in Section 4.1) with the highest number of judges selecting the correct baby (B), pictured wearing a hat. However, when the parent is pictured, the judges detect a resemblance between the parent and baby D, with the other three babies receiving about the same number of selections (as in Survey FS1).

Table 6.   Contingency Table of Baby Selection for Two Baby Selection Questions on Survey FS2

Baby Selected Without Parent Picture	Baby Selected With Parent Picture				Total
	A	*B*	*C*	*D*	
A	11	6	2	11	30
B	14	14	15	23	66
C	5	11	14	24	54
D	10	3	11	21	45
Total	40	34	42	79	195

5. Classroom Uses

This study and the general area of testing for resemblance has proven to be an interesting topic for students. Many students have an opinion about the resemblances between them and their family members. As a result, we have had many interesting discussions with our students about testing for resemblance between family members. Students are also interested to hear about research into resemblance for non-related family members and pets; such as the debate about whether dogs resemble their owners (Roy and Christenfeld 2004; Levine 2005; Roy and Christenfeld 2005) and evidence for resemblance between husbands and wives and between wives and their mother-in-laws (Bereczkei, Gyuris, P., Koves, and Bernath 2002).

We have used the material in Sections 2 through 4 of this article in a variety of courses, ranging from introductory to advanced. In the introductory statistics course, we use the material from Section 4.2 for one of the parent/child pairs as a lecture example. Students are then usually asked to analyze the data on the other parent/child pair on a homework assignment. In class, we have students discuss the appropriate graphical representations of the results (bar graph or pie graph, segmented bar graph or mosaic plot) along with the corresponding analysis. We make sure to discuss the research question, the data collection methods used in gathering the responses, and the limitations of our conclusions about the resemblance of the parent/child pair due to the use of a fixed set of foils. In this way, students see the entire research process (Franklin et al. 2005) from research question to data collection and analysis to conclusions even though they did not conduct the research themselves.

In higher level courses, the material we use from Section 4 depends on the topics covered in the course. In a course for in-service mathematics teachers, material from Sections 4.1 and 4.2 is presented as the corresponding topic is covered during lecture. Data from one

parent/child pair is used as a lecture example of the topic and data on the other parent/child pair is given as a homework assignment. In a course in applied categorical data analysis, the data from Sections 4.1, 4.2, and 4.4 is given as a mini course project where students are given the research questions from Section 2 and asked to analyze the data in order to answer them. Although not needed in order to answer the research questions, students will generally also report a confidence interval for the proportion from Section 4.2 and confidence intervals for the difference in two proportions from Section 4.1 and 4.4 for statistically significant results. As a result, several students presented analyses in their project reports similar to those found in Section 4.3 without any prompting from us.

Instructors can incorporate this study into their courses in several different ways. Instructors could use the data and analyses from this article directly according to the topic or topics presented in the course. However, before presenting the data and working through the analysis, we encourage instructors to discuss the research question and data collection procedures for each example. The results of the data analysis should then be used to answer the research question, keeping in mind the limitations of our conclusions due to the use of a fixed set of foils. Instructors could also collect data from their own students using our survey questions and pictures, which are available from the authors upon request. Data could be collected by paper and pencil, through a web-based course management system, through a web-based survey program, such as Survey Monkey, or by using personal responses systems (clickers) during class. Depending on class size, instructors may need to combine their data with ours or combine student responses from several semesters to obtain samples large enough to use the methods from Section 4.

Instructors and students could also conduct their own study of resemblance using our study as a template. Instructors could use our surveys and pictures and vary different aspects of the study design. For example, instructors might choose to vary the number of babies presented (e.g., three instead of four) or to randomly assign placement and labels to the baby pictures separately for each judge. Instead of using our pictures, instructors could decide to test for resemblance using a different parent/child pair. In this case, we suggest trying to enlist the help of colleagues and friends and family to obtain a large pool of foils to choose from. In the introductory course, instructors could collect the data from their students and analyze the data as a part of the same course. In higher level courses, instructors could work with their students to develop the research questions and study design. Students could then collect data for the study from students enrolled in other courses on campus, conduct the data analysis, and present their findings.

6. Conclusions

Like many new parents, we heard claims of a striking resemblance between us and our children. Our skepticism of these claims lead us to devise a study to objectively evaluate resemblance for specific parent/child pairs using us and our children as examples. We were right to be skeptical; neutral observers failed to detect a resemblance between the second author and his son. Although they did detect a resemblance between the first author and her daughter, based on the data, we cannot conclude that the first author looks more like her daughter than each of the other babies pictured.

Our intention in conducting this study was to use the surveys and resulting data as interesting examples in our teaching and to motivate methods for categorical data analysis. Our students have found this study and other studies on resemblance very interesting. In most cases, the study designs and analyses are easy for students to understand and can be used effectively in the classroom as motivation for studying topics in categorical data analysis.

Acknowledgments

References

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, Hoboken: Wiley.

Alvergne, A., Faurie, C., and Raymond, M. (2007), "Differential facial resemblance of young children to their parents: Who do children look like more?," *Evolution and Human Behavior*, 28, 135–144.

Bereczkei, T., Gyuris, P., Koves, P., and Bernath, L. (2002), "Homogamy, genetic similarity, and imprinting; parental influence on mate choice preferences," *Personality and Individual Differences*, 33, 677–690.

Christenfeld, N., and Hill, E.A. (1995), "Whose baby are you?," *Nature*, 378, 669.

DeVeaux, R.D., Velleman, P.F., and Bock, D.E. (2009), *Intro Stats*, Boston: Pearson, Addison Wesley.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2005), Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework, Alexandria, VA: American Statistical Association. Online: www.amstat.org/education/gaise/.

Levine, D. W. (2005), "Do Dogs Resemble Their Owners? A Reanalysis of Roy and Christenfeld (2004)," *Psychological Science*, 16, 83–84.

McNemar, Q. (1947), "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, 12, 153–157.

Moore, D.S., McCabe, G.P., and Craig, B. (2009), *Introduction to the Practice of Statistics*, New York: W.H. Freeman.

Nettleton, D. (2009), "Testing for the Supremacy of a Multinomial Cell Probability," *Journal of the American Statistical Association*, 104, 1052–1059.

Roy, M. M., and Christenfeld, N. J. S. (2004), "Do Dogs Resemble Their Owners?" *Psychological Science*, 15, 361–363.

Roy, M. M., and Christenfeld, N. J. S. (2005), "Dogs Still Do Resemble Their Owners," *Psychological Science*, 16, 743–744.

Stuart A. (1955), "A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification," *Biometrika*, 42, 412–416.

Amy G. Froelich
Department of Statistics
Iowa State University
3109 Snedecor Hall
Ames, IA 50011-1210
E-mail: amyf@iastate.edu

Dan Nettleton
Department of Statistics
Iowa State University

2115 Snedecor Hall
Ames, IA 50011-1210
E-mail: dnett@iastate.edu