Received: 7 August 2022 | Revised: 13 January 2024 | Accepted: 26 February 2024

DOI: 10.1111/desc.13499

Developmental Science WILEY

RESEARCH ARTICLE

Large-scale data decipher children's scale errors: A meta-analytic approach using the zero-inflated Poisson models

Hiromichi Hagihara[1,2] | Mikako Ishibashi[3] | Yusuke Moriguchi[4] | Yuta Shinya[5]

[1]Graduate School of Human Sciences, Osaka University, Suita, Osaka, Japan

[2]International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo Institutes for Advanced Study, Bunkyo, Tokyo, Japan

[3]Department of Psychology and Humanities, Edogawa University, Nagareyama, Chiba, Japan

[4]Graduate School of Letters, Kyoto University, Kyoto, Japan

[5]Graduate School of Education, The University of Tokyo, Bunkyo, Tokyo, Japan

Correspondence
Hiromichi Hagihara, Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita-shi, Osaka 565-0871, Japan.
Email: hiromichi.h@gmail.com

Funding information
JSPS KAKENHI, Grant/Award Numbers: JP18J21948, JP22KJ0525, JP22K13664; the Center for Early Childhood Development, Education, and Policy Research (Cedep); Graduate School of Education, The University of Tokyo

Abstract

Scale errors are intriguing phenomena in which a child tries to perform an object-specific action on a tiny object. Several viewpoints explaining the developmental mechanisms underlying scale errors exist; however, there is no unified account of how different factors interact and affect scale errors, and the statistical approaches used in the previous research do not adequately capture the structure of the data. By conducting a secondary analysis of aggregated datasets across nine different studies ($n = 528$) and using more appropriate statistical methods, this study provides a more accurate description of the development of scale errors. We implemented the zero-inflated Poisson (ZIP) regression that could directly handle the count data with a stack of zero observations and regarded developmental indices as continuous variables. The results suggested that the developmental trend of scale errors was well documented by an inverted U-shaped curve rather than a simple linear function, although nonlinearity captured different aspects of the scale errors between the laboratory and classroom data. We also found that repeated experiences with scale error tasks reduced the number of scale errors, whereas girls made more scale errors than boys. Furthermore, a model comparison approach revealed that predicate vocabulary size (e.g., adjectives or verbs), predicted developmental changes in scale errors better than noun vocabulary size, particularly in terms of the presence or absence of scale errors. The application of the ZIP model enables researchers to discern how different factors affect scale error production, thereby providing new insights into demystifying the mechanisms underlying these phenomena. A video abstract of this article can be viewed at https://youtu.be/1v1U6CjDZ1Q

KEYWORDS
Bayesian meta-analysis, count data, language development, scale error, toddlerhood, zero-inflated Poisson model (ZIP)

Research Highlights

- We fit a large dataset by aggregating the existing scale error data to the zero–inflated Poisson (ZIP) model.

- Scale errors peaked along the different developmental indices, but the underlying statistical structure differed between the in-lab and classroom datasets.
- Repeated experiences with scale error tasks and the children's gender affected the number of scale errors produced per session.
- Predicate vocabulary size (e.g., adjectives or verbs) better predicts developmental changes in scale errors than noun vocabulary size.

1 | INTRODUCTION

Scale errors are intriguing developmental phenomena which occur when a child tries to perform an object-specific action on a tiny object despite being impossible; for example, a child attempts to get into a miniature car (DeLoache et al., 2004). Children start to produce scale errors as early as approximately 12 months (Ware et al., 2010), increase their frequency during toddlerhood with a peak at around 18–24 months of age, and then become less likely to produce errors as they develop (DeLoache et al., 2004; Grzyb, Cangelosi, et al., 2019). Scale errors are robustly observed in various settings, including the classroom (Rosengren, Carmichael, et al., 2009; Rosengren et al., 2010) and at home (Rosengren, Gutiérrez, et al., 2009; Ware et al., 2010).

The mechanisms underlying scale errors have been explored from several viewpoints. Some scholars argue that scale errors are the result of children's immature inhibitory control, which cannot suppress inappropriate motor plans, such as getting inside a miniature-sized car (DeLoache et al., 2013; Ishibashi & Moriguchi, 2021; Rivière et al., 2020). Others attribute scale errors to children's developing size perception/comprehension of objects (Grzyb et al., 2017; Ishibashi & Moriguchi, 2017; Ware et al., 2006) or their own bodies (Brownell et al., 2007), or to the difficulty of integrating multiple visual features of objects, for example, local/global properties (Ishibashi et al., 2021).

In particular, the relationship between children's language development and scale errors has received considerable attention (Grzyb et al., 2014; Grzyb, Cangelosi, et al., 2019; Grzyb, Nagai, et al., 2019; Hagihara et al., 2022b; Hunley & Hahn, 2016; Oláh et al., 2016). Semantic conceptual systems that emerge through language development influence how children perceive objects, which further affects how they interact with objects. Increased attention to a particular feature of object shape (Gershkoff-Stowe & Smith, 2004; Landau et al., 1988; Smith et al., 2002) or object function (Kemler Nelson et al., 2000; Kobayashi, 1997; Zuniga-Montanez et al., 2021) helps a child learn word meanings. A strong bias for a certain type of object features develops and leads to the execution of object-specific actions regardless of other associated features, such as object size (Casler et al., 2011; Grzyb, Cangelosi, et al., 2019). In fact, it has been reported that children's vocabulary size, which reflects their language development as well as language-related attentional biases (e.g., Jones, 2003; Smith et al., 2002), is related to scale errors (Grzyb et al., 2014; Grzyb, Cangelosi, et al., 2019; Hagihara et al., 2022b). This finding suggests that children's language development may be a key factor in the mechanisms underlying scale errors.

Although various studies have investigated scale errors since the original study (DeLoache et al., 2004), at least three major concerns remain, leading to difficulties in understanding the developmental characteristics of scale errors: how and when they occur. First, there is no unified account of how various developmental and/or contextual factors influence scale error observations differently. As different cognitive and language abilities have been considered as factors related to scale errors, scale errors may result from a combination of different mechanisms. For instance, children with immature inhibitory control may repeatedly produce scale errors, whereas those with greater inhibitory control may produce few scale errors because they quickly switch their actions to a size-appropriate manner (Rosengren et al., 2010). Rapid developmental changes in children's semantic conceptual systems may drive the occurrence of scale errors (Grzyb, Nagai, et al., 2019; Hagihara et al., 2022b). Everyday contexts with a longer observation time may increase opportunities to observe scale errors, whereas laboratory settings may underestimate how often children produce scale errors (Rosengren et al., 2009; Rosengren et al., 2010). Disentangling which developmental/contextual factors affect scale errors differently, such as repeated attempts once a scale error is produced, or whether scale errors occur or not, will lead to a deeper understanding of the mechanisms of how scale errors occur.

Second, most studies on scale errors regarded children's age as a discrete variable by classifying participants into several age groups (Hagihara et al., 2022b). Such arbitrary categorizations of continuous variables can statistically distort the research findings (Naggara et al., 2011; Royston et al., 2006; Rucker et al., 2015). Some studies have stated that the developmental trend of scale errors can be drawn as an inverted U-shaped curve with a peak at around 21–24 months (e.g., DeLoache et al., 2004), whilst other studies showed that scale error occurrence can be sufficiently documented just by a linear decrease function during toddlerhood with a peak at around 18–20 months (e.g., Grzyb et al., 2019). Precise detection of when scale errors are most likely to occur is important because it helps specify the relevant cognitive/language ability. For instance, a strong attentional bias toward object shape can emerge at 17–19 months (Smith et al., 2002), or this shape bias noticeably develops when children's productive vocabulary size is between 51 and 100 words (Gershkoff-Stowe & Smith, 2004). If the probability of scale error occurrence peaks after these reported

shape bias milestones, one can assume that shape bias may cause scale errors; however, if this relationship turns out to be a flipped order, then scale errors are assumed to occur independently of shape bias.

Third, the statistical methods used in previous studies were inadequate, given the structure of the scale error data, which are generally less frequent events. The original study observed that approximately 46.3% of children aged 18–30 months produced at least one scale error (DeLoache et al., 2004). This indicates that the distribution of the number of scale errors is considerably skewed toward zero, violating the normality assumption. Nevertheless, most empirical studies make statistical inferences based on ordinary linear regression, which supports this assumption. The distribution of scale errors that contain many zeros raises a further problem: a count of zero sometimes arises from more than one way (McElreath, 2020). When observing a value of zero in a scale error task, a researcher cannot distinguish whether the child does not produce scale errors at all, or whether the child who produces scale errors does not execute them at the time of observation.

To set the stage for future research directly investigating the relationships between scale errors and other cognitive/language abilities and providing a unified account of the mechanisms underlying scale errors, this study aimed to address the above-mentioned concerns by conducting a secondary analysis of aggregated datasets across nine different studies with an appropriate statistical approach. Integrating different datasets, which are usually applied in meta-analyses, facilitates robust conclusions regarding developmental phenomena and characterizes developmental changes (Bergmann et al., 2018; Lewis et al., 2020). Merging a small number of datasets is still beneficial for detecting effects with small sizes (Goh et al., 2016). Furthermore, a large dataset allows us to perform complex statistical modeling with many independent variables because accurate estimates generally require a larger sample size as the number of predictors increases (Austin & Steyerberg, 2015; Bujang et al., 2018).

To adequately capture the structure of the scale error data and their developmental trends, we adopted two statistical strategies. First, we regarded children's age in months and vocabulary size as a continuous variable and constructed statistical models so that a nonlinear developmental trend in scale errors (e.g., inverted U-shaped curve) can be expressed. Second, we utilized a statistical model called zero–inflated Poisson (ZIP) regression (Lambert, 1992) that can directly cope with the count data with a stack of zeros, allowing for a better understanding and estimation of the occurrence of scale errors. The ZIP model has been used in various fields of research, including psychology, medicine, and ecology (e.g., Atkins & Gallop, 2007; Böhning et al., 1999; Hu et al., 2011; Karazsia & van Dulmen, 2008; Loeys et al., 2012; Martin et al., 2005; Wiesner & Kim, 2006). The ZIP model allows us to simultaneously and discernably estimate the probability of the occurrence of a scale error at a given degree of development and the probability of how many times scale errors are observed if they occur. Thus, it is possible to examine separately from this single model whether developmental linearity or nonlinearity concerns the presence or absence of scale errors (i.e., logistic part), the number of scale errors observed (i.e., count part), or both.

Taking advantage of an aggregated large dataset and an adequate statistical approach, this study examined how scale errors could be described as a function of age in months using laboratory and classroom data (Analysis 1). We also used partial data to investigate how scale errors could be described as a function of another developmental index reflecting children's language abilities and vocabulary size (Analysis 2). We further compared the subvocabulary sizes of nouns, verbs, and adjectives that would better predict scale errors to narrow down the possible hypotheses about the mechanisms underlying scale errors.

2 | DESCRIPTION OF THE AGGREGATED DATASETS

2.1 | Data collection strategy

Because we needed the raw data to perform a secondary analysis, we obtained them from three different sources: the existing datasets in hands, the call for providing the raw data via mailing lists, and contacts to scale error researchers who were assumed to have relatively large data on children's scale errors. We reached developmental scientists internationally using two mailing lists from the International Congress of Infant Studies (ICIS) and the Cognitive Development Society (CDS) from April to May 2023. We explicitly mentioned that both published and unpublished works are welcomed.

2.2 | Participant characteristics

We obtained 766 data points from 528 children (gender: 228 girls, 38 unknown; age: 13–41 months, $M = 22.6$, $SD = 4.3$) from nine studies. Among the data points, 355 (47.0%) produced at least one scale error. This proportion was similar to that in the original study in which 46.3% of children aged 18–30 months produced at least one scale error (DeLoache et al., 2004). Thus, this merged dataset was considered a representative dataset of scale errors. This large dataset included 439 data points from in-lab experimental settings[1] (Table 1) and 317 data points from classroom observational settings (Table 2). All the in-lab data were cross-sectional, whereas all classroom data were longitudinal.

Rosengren et al. (2009) observed 68 children in six classrooms and recorded their interactions with the target toys. The provided dataset contained 38 children who were observed performing scale errors (55.9%), indicating that the other 30 children did not play with the target toys during the observations[2]. Note that information about the children's gender was not included in this dataset, although it was reported that there were equal numbers of boys and girls in each classroom (Rosengren et al., 2009). Similarly, Rosengren et al. (2010) observed 24 children divided into two classrooms; however, their raw data included 21 children who were observed performing scale errors (87.5%)[3]. The dataset of Rosengren (n.d.) was obtained from an unpublished study.

TABLE 1 Existing datasets with in-lab settings used in this study.

Article	*n*	%Girls	Age		Country	# Objects	%Scale errors	# Scale errors	
			Range	*M* (*SD*)				Range	*M* (*SD*)
Arterberry et al. (2020)	67	38.8	23–30	25.7 (2.3)	US	3	44.8	0–7	1.0 (1.6)
Grzyb, Cangelosi et al. (2019)	125	44.0	17–29	23.0 (3.0)	UK	3	38.4	0–7	0.8 (1.4)
Hagihara et al. (2022b)[a]	72	59.7	18–31	23.4 (3.9)	JP	5	38.9	0–5	0.8 (1.3)
Ishibashi and Moriguchi (2017)	54	40.7	16–37	24.1 (5.4)	JP	4	48.1	0–7	1.4 (1.9)
Ishibashi et al. (2021), Japan sample	40	42.5	18–24	19.7 (1.6)	JP	4	47.5	0–6	1.3 (1.7)
Ishibashi et al. (2021), UK sample	40	52.5	18–23	19.8 (1.8)	UK	3	37.5	0–7	0.9 (1.6)
Ishibashi and Uehara (2020)[b]	41	43.9	15–35	23.5 (6.0)	JP	4	56.1	0–6	1.3 (1.6)
Total	439	46.0	15–37	23.1 (4.1)			43.1	0–7	1.0 (1.5)

[a]Data from Hagihara et al. (2022b) were published in Hagihara et al. (2022a).
[b]The participants in Ishibashi and Uehara (2020) partially overlapped with those in Ishibashi et al. (2021). Participants' information, excluding the overlap, is provided here.

TABLE 2 Existing datasets with classroom settings used in this study.

Article	# Children	%Girls	# Data points	# Data points per child		Age		Country	# Objects	%Scale errors	# Scale errors	
				Range	*M* (*SD*)	Range	*M* (*SD*)				Range	*M* (*SD*)
Rosengren et al. (2009)[a]	38	NA	60	1–4	1.6 (0.8)	13–41	24.5 (7.7)	US	3	78.3	0–11	2.0 (2.1)
Rosengren et al. (2010)	21	52.4	75	1–9	3.6 (2.2)	18–30	23.5 (2.8)	US	4	88.0	0–12	1.8 (2.1)
Rosengren (n.d.)[b]	30	50.0	182	1–12	6.1 (3.4)	13–27	20.7 (3.4)	US	3	29.1	0–6	0.6 (1.2)
Total	89	51.0[c]	327	1–12	3.7 (3.2)	13–41	22.0 (4.6)			51.1	0–12	1.1 (1.7)

[a]Children's gender was not included in the dataset provided from Rosengren et al. (2009).
[b]Unpublished data.
[c]Calculated from Rosengren et al. (2010) and Rosengren (n.d.).

2.3 | Scale error task

The scale error tasks in each study were based on DeLoache et al. (2004). Most of the in-lab tasks used a car, chair (often with a table), and slide with two different sizes: child-sized and corresponding miniature-sized objects. Exceptions for experimental materials were described in Supplementary Description 1. For scale error tasks in classroom settings, Rosengren et al. (2009) used one of three sets of miniature-sized toys consisting of three toys each: Set A included a car, a slide, and a sofa; Set B included a rocking chair, a Hummer vehicle, and a bed; and Set C included a bathtub, a car, and a wagon. Rosengren et al. (2010) used four items: a couch, slide, bed, and car. Rosengren et al. (n.d.) used a slide, bed, and car. None of the classroom studies used child-sized (i.e., bigger-sized) objects; however, there could have possibly been child-sized objects available outside of the classroom (e.g., outdoor play areas).

In most of the in-lab experiments, each participant freely played with the child-sized objects for approximately 5–7 min in the playroom. If the participant showed little interest in the objects, the experimenter drew attention to them and encouraged the children to interact with them. Subsequently, the participant left the playroom temporarily and the experimenter replaced the objects with miniature-sized objects. The child then returned to the playroom and played with the replaced objects for another 5–7 min. According to Arterberry et al. (2020), children only experienced the latter interaction phase.

In Hagihara et al. (2022b), children performed a scale error task twice with a mean interval of 13.1 days (*SD* = 6.0) in one session where specific object labels were provided (the noun condition; e.g., "Look at the car!") and the other session where only general pronouns were provided (the pronoun condition; e.g., "Look at this!"). The order of the labeling conditions was counterbalanced. Since the experimenter's verbal instructions are not restricted during a scale error task in general, we included only the dataset of the noun condition in the present study. Ishibashi and Uehara (2020) analyzed only the first 3 min of the whole 5 min observation period; however, to align the time of analysis with other studies, we recalculated the number

of scale errors for all 5-min observation periods and used them in this study.

For the classroom studies, observations were performed from observation booths adjacent to each classroom equipped with one-way mirrors and earphones. An experimenter placed the target toys in the classroom clearly observable from the observation booth before each observation and the children freely played with them in the classroom, which was also filled with other classroom toys and materials. There were 8–14 children in each classroom (Rosengren et al., 2009, 2010). The target toys were removed after each observation session. Rosengren et al. (2009) conducted an average of 70 min of observations over a 3-month period, whereas Rosengren et al. (2010) generally performed 20-min observation sessions, each over a 10-week period. The observation period for Rosengren et al. (n.d.) was 2 months; however, the duration of each observation session was not provided because of unpublished work.

2.4 | Coding scheme

In general, the participants' actions on the miniature-sized objects were coded as scale errors based on the coding scheme given by DeLoache et al. (2004): (a) whether the participant tried to interact with objects in the same way they performed on the child-sized objects, (b) whether the participant's body part(s) touched the objects' adequate part(s), and (c) whether the participant's attempt to wards the objects was serious. For criterion (c), a 5-point Likert scale ranging from 1 (definitely serious) to 5 (definitely pretending) was used, and the action was classified as a scale error when it was scored as 1 or 2. The inter-rater reliability was confirmed in each study (see Supplementary Description 1).

For the classroom dataset, we limited our focus to body-based scale errors to make the laboratory and classroom data as equivalent as possible. For instance, if a child attempted to put a doll into a tiny toy car despite being impossible (object-based scale errors), we did not count it as a scale error. To count the number of scale errors in these studies, we conservatively converted the original coding into an integer variable, where the categorical coding of the number of attempts of 1, 2, 3–4, and >4 were converted into 1, 2, 3, and 5, respectively. The data points for the same child and date were collapsed and merged. Therefore, each data point in our final dataset represented whether and how many children produced scale errors per observation day.

3 | ANALYSIS 1: AGE-RELATED CHANGE

3.1 | Analysis approach

3.1.1 | Developmental indices

We used children's age in months as an independent variable. To allow the models to express a nonlinear age-related change in scale errors, two types of regression were implemented: simple linear and quadratic functions (Figure S1).

3.1.2 | Zero-inflated Poisson model

The ZIP regression is a model used to fit the count data with a high incidence of zeros (Lambert, 1992). This model assumes that the zero observations derive from two different processes: "structural" and "sampling" (Hu et al., 2011). Let us say you tried to catch a particular species of fish in a river, and you failed, which means zero observation. In this case, you might have failed because the river was not a habitat for that species of fish in the first place, which is called structural zeros. Meanwhile, it may have been the case that the fish species were present in the river, but you failed to catch them by chance, which scores sampling zeros. The ZIP model explicitly discerns and then mixes these two sources of zeros by assuming a Bernoulli distribution model for the structural zero observations and a Poisson distribution for the count part, including the sampling zeros (Figure 1). Statistical analyses using the ZIP model are seen in a variety of research fields, including psychology, medicine, and ecology (e.g., Atkins & Gallop, 2007; Böhning et al., 1999; Hu et al., 2011; Karazsia & van Dulmen, 2008; Loeys et al., 2012; Martin et al., 2005; Wiesner & Kim, 2006).

In our case, the number of scale errors observed y can be formulated as

$$\mathrm{ZIP}(y|\theta,\ \lambda) = \begin{cases} (1-\theta) + \theta \times Poisson(0|\lambda), & if\ y = 0 \\ \theta \times Poisson(y|\lambda), & if\ y \geq 1, \end{cases}$$

where θ from a Bernoulli distribution is the probability of producing scale errors, whereas λ from a Poisson distribution is the probability of the number of scale errors produced in a session if they are observed at least once.

3.1.3 | Model implementation and model comparison

We fit the scale error data using the ZIP regression with a logit link function for the structural zero part (θ) and a log link function for the sampling zero part (λ). We used Bayesian generalized linear models using CmdStanR 0.5.3 (Gabry & Češnovar, 2020), an interface to Stan (Stan Development Team, 2021), for model fitting. The number of scale errors observed was regarded as the dependent variable and age in months was regarded as the independent variable. To avoid constructing complex models that are difficult to converge, we did not include random effects. Instead, we included possible covariates (i.e., country, session, number of target objects, task duration, and gender) for both the structural and sampling zero parts (see Supplementary Description 2 for details).

The model candidates are listed in Table 3. For instance, Model 1 included quadratic functions of age for both the Bernoulli and Poisson regressions, whereas Model 2 included the term of age for Bernoulli regression (i.e., simple linear function) and age and squared age for Poisson regression (i.e., quadratic function). For reference, we constructed Models 5 and 6, that performed a simple Poisson regression using linear and quadratic functions.

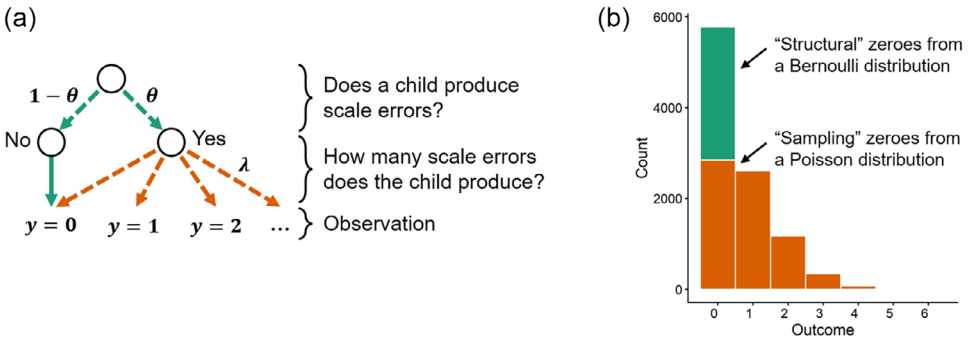

FIGURE 1 A schematic view of the zero-inflated Poisson model. (a) Calculation flow of the zero-inflated model in the case of scale error production. (b) Histogram of zero-inflated observations generated from the simulation data ($\theta = 0.7, \lambda = 0.9$). The "structural zeros" derived from the Bernoulli distribution increase zero observations in total; however, in real data, a researcher cannot discern, which zeros come from which process (McElreath, 2020).

TABLE 3 Model candidates and widely applicable information criterion (WAIC) values in Analysis 1.

			WAIC (SE)	
Model number	Distribution	Functions of age for regression	In-lab	Classroom
Model 1	ZIP	θ, λ: Quadratic	1206.50 (46.54)	1006.05 (57.13)
Model 2	ZIP	θ: Linear; λ: Quadratic	1206.68 (46.52)	**1002.78 (56.89)**
Model 3	ZIP	θ: Quadratic; λ: Linear	**1204.26 (46.29)**	1006.69 (58.90)
Model 4	ZIP	θ, λ: Linear	1204.69 (46.25)	1003.91 (58.57)
Model 5	Poisson	θ: NA; λ: Quadratic	1384.99 (58.51)	1039.53 (67.88)
Model 6	Poisson	θ: NA; λ: Linear	1389.96 (57.84)	1038.00 (68.89)

Note: For both θ and λ regression equations, covariates were also included (see the main text and Supplementary Description 2 for details). The WAIC values for the best models among the model candidates are shown in boldface.

For each fixed effect, we used weakly informative Student's t priors ($\nu = 3, \mu = 0, \sigma = 1$), whose parameters were determined based on a previous study (Hagihara et al., 2022b), to stabilize the parameter estimates. We set four chains and iterations of 15,000 with warm-up samples of 2000. We confirmed whether the R-hat values were below 1.1 (Gelman et al., 2013) to verify the convergence of parameter estimates. For interpretation, we used the posterior median (MED) and 95% Bayesian credible intervals (CIs) for the parameter estimates or expected values. We used a model comparison approach using the widely applicable information criterion (WAIC) (Watanabe, 2010a, 2010b) to examine which model best predicted scale errors. The smaller the WAIC, the better the model.

3.2 | Results

3.2.1 | In-lab data

Based on a model comparison approach using the WAIC values, the best model included age and squared age for the Bernoulli regression, but did not include the squared age for the Poisson regression of the ZIP model (i.e., Model 3, WAIC = 1204.26, *SE* = 46.29; Table 3). This suggests that scale errors are a curved function of age in terms of whether a child produces them or not, whereas they can be drawn as a simple linear function of age when it comes to how many scale errors are produced. The models using a mere Poisson regression (Models 5 and 6) were much inferior to the models using the ZIP regression, suggesting that the stacked zero should be considered for scale errors. To further assess the adequacy of the model fit for the best model, we performed a posterior predictive check. This revealed that 98.2% of the data points fell within the predicted 95% CI of [0, 5], demonstrating adequate model fit.

Although the best model included the term Age^2 for the Bernoulli regression, this parameter estimate straddled zero (MED = −0.16 [−0.39, 0.05]; Table 4), whereas the term Age had a negative effect (MED = −0.40, CI[−0.70, −0.13]). Thus, it was evident that the probability of producing scale errors decreased as a function of age within the window of 15−37 months. We also detected the effects of sessions (MED = −1.06, CI[−1.77, −0.36]) and gender (MED = 0.32, CI[0.10, 0.55]) on the count part of scale errors. Repeated engagement in a scale error task led to a decreased number of errors, whereas girls were more likely to produce scale errors repeatedly than boys.

The developmental trend of the expected values of θ, λ, and y is shown in Figure 2. We calculated the expected values while country and gender were averaged, and the session, number of objects, and task duration were specified as the 1, 3, and 5 min, respectively.

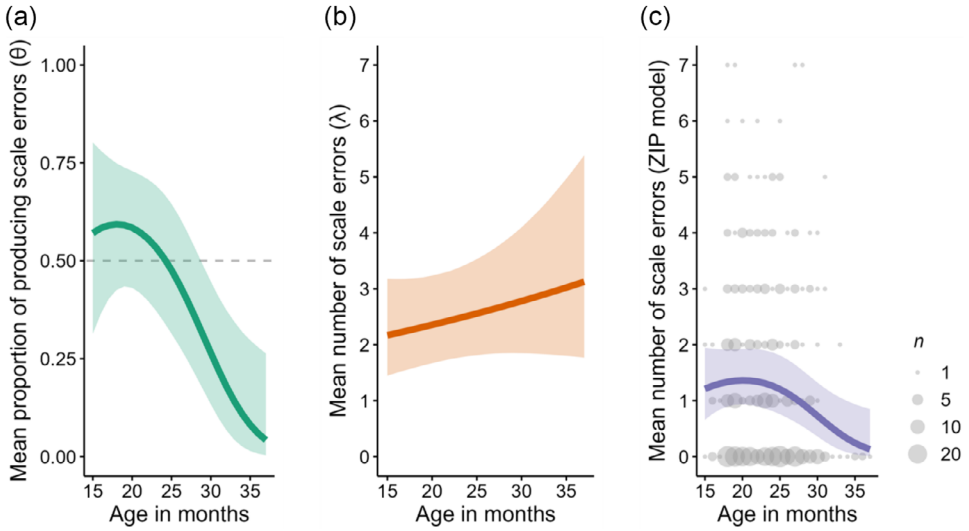

FIGURE 2 Developmental changes in scale errors drawn from the best model in Analysis 1 (in-lab data, Model 3). Developmental trend for the proportion part (a), count part (b), and overall zero-inflated Poisson (c). The thick lines and shaded areas indicate the posterior median and 95% Bayesian credible intervals of the expected values from the selected best model. The covariates of country and gender were averaged for all graphs, and those of the session, the number of objects, and task duration were specified as the 1, 3, and 5 min, respectively. For (a), the dashed horizontal line indicates chance level. In (c), the size of the data points represents the number of children in the same coordinates.

When looking at the expected value of θ, the probability of producing scale errors peaked at 18 months of age (MED = 0.59, CI[0.43, 0.75]) and then decreased with age (MED = 0.04, CI [0.00, 0.26] at 37 months). However, the difference in the expected value of θ between 18 and 15 months still straddled zero (MED_{18-15} = 0.02, CI[−0.11, 0.17]), although a difference between 18 and 37 months was detected (MED_{18-37} = 0.53, CI[0.33, 0.68]). For the expected value of λ, the mean number of scale errors produced in the first session did not show a clear developmental trend from 15 months (MED = 2.12, CI [1.44, 3.18]) to 37 months (MED = 3.13, CI [1.77, 5.38]). When mixing both proportion and count parts, the mean number of scale errors observed, peaked at 20 months of age (MED = 1.36, CI [0.93, 1.93]). The MEDs for 15 and 37 months were 1.21 (CI [0.66, 1.94]) and 0.13 (CI [0.01, 0.85]), respectively. As in the expected value of θ, the difference in the expected number of scale errors between 20 and 15 months included zero ($MED_{20-15\ months}$ = 0.14, CI[−0.34, 0.64]) whereas that between 20 and 37 months exceeded zero ($MED_{20-37\ months}$ = 1.18, CI[0.46, 1.79]), suggesting a clear decrease of scale errors as a function of age.

We finally examined the correlation coefficient between θ and λ to determine whether the higher the probability of scale errors was at least once, the more times they were observed. Spearman's rank correlation coefficient did not show any specific direction (MED = −0.22, CI[−0.66, 0.32]), suggesting that there were no associations between whether the scale errors were produced and how many scale errors were produced when observed.

3.2.2 | Classroom data

A model comparison approach revealed that the best model included age but not squared age for the Bernoulli regression, whereas it included both age and squared age terms for the Poisson regression (Model 2, WAIC = 1002.78, *SE* = 56.89; Table 2). This suggests that, in contrast to the in-lab data, scale errors observed in classroom settings are drawn as a simple linear function of age in terms of the proportion part, although they are a curved function of age for the count part. The ZIP regression was superior to the Poisson regression models (Models 5 and 6). A posterior predictive check demonstrated that 95.0% of the observations were within the 95% CI of the posterior predictive distribution, [0, 4], suggesting model fitting adequacy.

Despite the selection of the best model, its parameter estimates for Age and Age^2 straddled zero for both the Bernoulli and Poisson regressions (Table 5). All the covariates included zero in the parameter estimates. The developmental change of the expected values of θ, λ, and y was calculated while the session and the number of target objects were specified as the first and three, respectively (Figure 3). In the 13–41-month age window, these expected values peaked at 41 months for θ (MED = 0.83, CI[0.51, 0.97]), 26 months for λ (MED = 2.06, CI[1.56, 2.69]), and 27 months for y (MED = 1.54, CI[1.19, 1.97]). However, no clear developmental change was detected, as the differences in the expected value between these peaks and youngest/oldest age included zero for θ ($MED_{41-13\ months}$ = 0.18, CI[−0.29, 0.51]), λ ($MED_{26-13\ months}$ = 0.82, CI[−0.17, 1.71]; $MED_{26-41\ months}$ = 1.01, CI[−0.60, 1.98]), and y ($MED_{27-41\ months}$ = 0.71, CI[−0.52, 1.42]), except for the difference between 27 and 13 months for y ($MED_{27-13\ months}$ = 0.75, CI[0.10, 1.35]). This suggests that although each component of the ZIP model might not have a clear developmental change owing to high dispersion, the observable events resulting from their combination appear to have a slightly clearer age-related change, which is an increase in scale errors from infancy to toddlerhood.

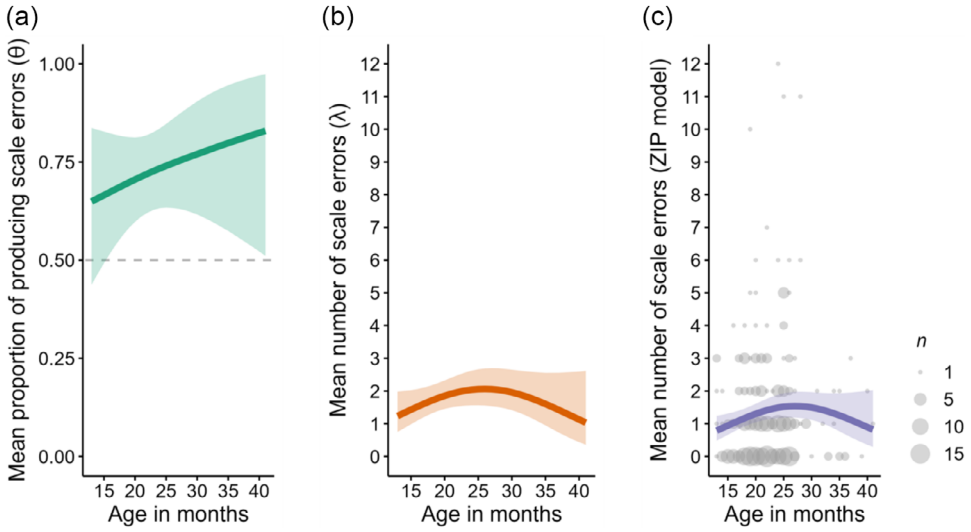

FIGURE 3 Developmental changes in scale errors drawn from the best model in Analysis 1 (classroom data, Model 2). Developmental trend for the proportion part (a), count part (b), and overall zero-inflated Poisson (c). The covariates of the session and the number of target objects were specified as the first and three. Other legends are the same as those in Figure 2.

TABLE 4 Posterior median and 95% Bayesian credible intervals (CIs) of parameter estimates for the best model of the in-lab data in Analysis 1 (Model 3).

Parameter	Posterior median [95% CI]
Logistic part of the model (θ)	
Intercept	0.39 [−0.48, 1.32]
Age	**−0.40 [−0.70, −0.13]**
Age^2	−0.16 [−0.39, 0.05]
Country (UK)	−0.73 [−1.74, 0.19]
Country (US)	−0.07 [−1.17, 1.02]
Session	0.57 [−0.73, 3.25]
# Objects	−0.02 [−0.68, 0.67]
Task duration	0.23 [−0.54, 1.01]
Gender (girl)	0.02 [−0.46, 0.49]
Count part of the model (λ)	
Intercept	0.90 [0.46, 1.36]
Age	0.07 [−0.06, 0.19]
Country (UK)	−0.24 [−0.80, 0.29]
Country (US)	−0.24 [−0.87, 0.38]
Session	**−1.06 [−1.77, −0.36]**
# Objects	−0.18 [−0.57, 0.17]
Task duration	−0.20 [−0.61, 0.24]
Gender (girl)	**0.32 [0.10, 0.55]**

Note: Children's age in months was standardized. The detected effects are indicated in bold. See Supplementary Description 2 for details about covariates.

As in the laboratory data analysis, there was no association between θ and λ based on their correlation coefficient (Spearman's rank correlation, MED = 0.60, CI[−0.54, 0.93]).

TABLE 5 Posterior median and 95% Bayesian credible intervals (CIs) of parameter estimates for the best model of the classroom data in Analysis 1 (Model 2).

Parameter	Posterior median [95% CI]
Logistic part of the model (θ)	
Intercept	0.95 [0.50, 1.51]
Age	0.16 [−0.23, 0.60]
Session	−0.50 [−1.07, 0.20]
Count part of the model (λ)	
Intercept	0.68 [0.43, 0.91]
Age	0.11 [−0.05, 0.28]
Age^2	−0.07 [−0.15, 0.01]
Session	−0.25 [−0.53, 0.02]
# Objects	0.09 [−0.18, 0.36]

Note: Children's age in months and sessions were standardized. See Supplementary Description 2 for details about covariates.

3.3 | Discussion

The model comparison approach demonstrated that compared with the models that simply used Poisson regression, the ZIP-model-based regression showed better predictability for both the laboratory and classroom datasets. Since scale errors are distributed with a stack of zero observations, it is important to analyze the scale error data using a statistical model that can cope with the characteristics of the data. Usage of the ZIP model is also useful for understanding the mechanisms underlying scale errors because it can discern two sources of zeros: the probability of producing scale errors (i.e., "structural zeros") and the number of scale errors (i.e., "sampling zeros"). For the in-lab data, the age effect was detected in the former part, whereas the number of sessions and children's gender contributed to the latter part.

We would speculate that what changes developmentally is whether the child produces scale errors and that other factors not directly related to development determine how many times the child produces them. The negative effect of repeated exposure to miniature objects was compatible with the results of an in-lab study (Hagihara et al., 2022b) and a classroom study (Rosengren et al., 2010). Previous memories of failing to execute object-specific actions may suppress subsequent scale errors, or children may simply lose interest in miniature-sized objects (Hagihara et al., 2022b). A prospective parental diary study reported that girls produced more scale errors than boys (Rosengren, Gutiérrez et al., 2009) was supported by the large laboratory dataset used in this study. Our results suggested that the difference in the number of scale errors between girls and boys was 0.35 on average. This gender difference could be attributed to girls finding small toys more engaging and interesting (Rosengren, Gutiérrez et al., 2009), or other developmental factors reported to differ by gender, such as lexical skills (Eriksson et al., 2012).

Age-related changes in scale errors were well documented by an inverted U-shaped curve rather than a simple linear function for both the laboratory and classroom data, although nonlinearity belonged to different sources of the two distributions and peaked at different ages. The developmental trend of scale errors collected in labs was elicited by an inverted U-shaped curve with a peak age at 18 months for the proportion part, whereas in terms of the count part, children's age in months was less relevant to the developmental change in scale errors. It is evident that the scale errors decrease as a function of age in toddlerhood, ranging from 15 to 37 months. In contrast, the model comparison for the classroom data showed that the proportion part was less related to age, but the count part could be described with a quadratic function of age, with a peak at 26 months.

What brought the differences between the in-lab experimental and naturalistic observational data? For the proportion part, different task settings in the classroom data may have resulted in a higher probability of detecting at least one scale error than in the laboratory data, making the nonlinear age effect less detectable. A longer observation duration in classroom settings than in-lab settings may have increased the possibility of observing scale errors. Because children were observed in environments familiar to them in the classroom data, they may have been relaxed and thus easily executed any actions that came to mind, leading to many scale errors. Given that the children could observe other children producing scale errors in classroom settings, exposure to other children's actions may have primed the participants to produce similar actions. For the count part, although the observed actions were the same, the developmental mechanisms underlying scale errors could differ between laboratory and classroom settings. Given that classroom data peak at 26–27 months, which is much later than in-lab data at 18–20 months, scale errors observed in classrooms may reflect the development of higher cognitive functions, such as the decontextualization of object concepts (Bigham & Bourchier-Sutton, 2007; Elder & Pederson, 1978). The combined use of a scale-error task and other developmental tasks related to candidate abilities will contribute to disentangling how scale errors differ between the two settings.

Although we selected relatively better models describing how scale errors can be depicted as a function of age, age itself might not be a good developmental predictor of scale errors, given that most 95% CIs were still wide and straddled zero. In Analysis 2, using the partial dataset of in-lab scale errors, we investigated how scale errors can be described as a function of vocabulary size, which reflects children's language abilities and is considered a relevant developmental predictor in scale error research (e.g., Grzyb, Cangelosi, et al., 2019; Hagihara et al., 2022b).

4 | ANALYSIS 2: VOCABULARY-RELATED CHANGE

4.1 | Datasets

From the datasets we used in Analysis 1, we extracted partial in-lab data that included children's productive vocabulary size, comprising 271 toddlers aged 15–35 months (131 girls; M_{age} = 22.8, SD = 3.9; Table 6)[4]. Children's productive vocabulary size was assessed using the Words and Grammar form of the Japanese MacArthur–Bates Communicative Development Inventory (J-MCDI; Ogura & Watamaki, 2004) for the Japanese sample and the Oxford Communicative Development Inventory (O-CDI; Hamilton et al., 2000) for the UK sample. The J-MCDI includes 711 vocabulary items and is used with Japanese-speaking children aged 16 and 36 months, whereas the O-CDI includes 418 items and is typically used with British English-speaking children aged from approximately 11–26 months.

Among the existing data, Grzyb, Cangelosi, et al. (2019) and Hagihara et al. (2022b) provided raw inventory data, enabling us to calculate the subvocabulary size of each part of speech separately: the vocabulary size of nouns, verbs, and adjectives. Based on the classification criteria used in previous studies (Caselli et al., 1999; Ogura et al., 2016), we calculated the children's subvocabulary size[5].

4.2 | Analysis approach

First, we used children's total productive vocabulary size as the developmental index using all data, including this measure (n = 271). For each developmental index, simple linear and quadratic functions were used as in Analysis 1. The ZIP model was used for the regression analysis, given the results obtained in Analysis 1. As a reference, we also constructed a ZIP model predicting scale errors from children's ages with the selected combination of age terms as best in Analysis 1. We included the covariates of country, session, number of target objects, task duration, and gender. Consequently, we constructed five model candidates (Table 7)[6]. Other settings were identical to those in Analysis 1.

Thereafter, we applied the same statistical modeling to the data containing the subvocabulary sizes of nouns, verbs, and adjectives (n = 197). We compared the models while changing the independent variables to detect the developmental index most sensitive to

TABLE 6 Existing datasets used in Analysis 2.

Article	*n*	Total vocabulary size		Nouns		Verbs		Adjectives	
		Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)
Grzyb, Cangelosi et al. (2019)	125	5–417	190.0 (132.0)	0–183	91.8 (60.0)	0–70	26.6 (25.5)	0–39	14.4 (13.3)
Hagihara et al. (2022b)	72	11–654	201.0 (162.0)	1–258	84.3 (71.7)	0–102	22.5 (28.5)	0–62	17.3 (16.8)
Ishibashi et al. (2021), Japan sample	40	2–238	85.0 (68.3)						
Ishibashi and Uehara (2020)[a]	34	0–700	262.0 (231.0)						
Total	271	0–700	186.5 (156.2)	0–258	89.1 (64.4)	0–102	25.1 (26.7)	0–62	15.5 (14.7)

[a]The sample size from Ishibashi and Uehara (2020) was smaller than that in Study 1 because, for some children, vocabulary size was not assessed ($n = 4$), or a different version of the inventory was performed ($n = 3$).

TABLE 7 Model candidates and widely applicable information criterion (WAIC) values in Analysis 2 (total vocabulary size).

Model number	Developmental index	Functions of developmental index for regression	WAIC (SE)
Model 1	Vocabulary size	θ, λ: Quadratic	711.00 (36.50)
Model 2	Vocabulary size	θ: Linear; λ: Quadratic	713.88 (36.16)
Model 3	Vocabulary size	θ: Quadratic; λ: Linear	**708.79 (36.18)**
Model 4	Vocabulary size	θ, λ: Linear	710.95 (35.71)
Model 5	Age (For reference)	θ: Quadratic; λ: Linear	708.83 (36.07)

Note: For both θ and λ regression equations, covariates were also included. The WAIC values for the best models among the model candidates are shown in boldface.

TABLE 8 Model candidates and widely applicable information criterion (WAIC) values in Analysis 2.

Model number	Functions of developmental index for regression	WAIC (SE)			
		Nouns	Verbs	Adjectives	Total vocabulary size (For reference)
Model 1	θ, λ: Quadratic	499.15 (31.31)	493.33 (31.47)	492.86 (31.36)	
Model 2	θ: Linear; λ: Quadratic	497.20 (31.24)	495.02 (31.34)	492.55 (31.35)	
Model 3	θ: Quadratic; λ: Linear	496.58 (30.94)	491.94 (31.26)	**490.82 (31.11)**	492.28 (31.05)
Model 4	θ, λ: Linear	494.86 (30.88)	493.01 (31.06)	492.17 (31.05)	

Note: For both θ and λ regression equations, covariates were also included. The WAIC values for the best models among the model candidates are shown in boldface.

scale errors (Table 8). In addition to each subvocabulary size, the best model using the total vocabulary size as an independent variable was constructed again as a reference.

4.3 | Results

4.3.1 | Total vocabulary size as the developmental index

Among the candidate models regarding total vocabulary size as a developmental measure, the best model was Model 3, in which a quadratic function of total vocabulary size was included in the Bernoulli regression, while its simple linear function was included in the Poisson regression (WAIC = 708.79, *SE* = 36.18; Table 7). The WAIC value for the best model was similar to that of the reference model, treating children's age as the developmental index (WAIC = 708.83, *SE* = 36.07), suggesting that the total vocabulary size was an equivalent predictor of age for describing the developmental trend of scale errors. A major difference in the best model from the findings in Analysis 1 was that we detected a clear inverted U-shaped curve of the developmental trend for the proportion part (Vocabulary size^2: MED = −0.36, CI[−0.76, −0.02]; Table 9). For the other fixed effects, we did not detect the effect of gender for the count part

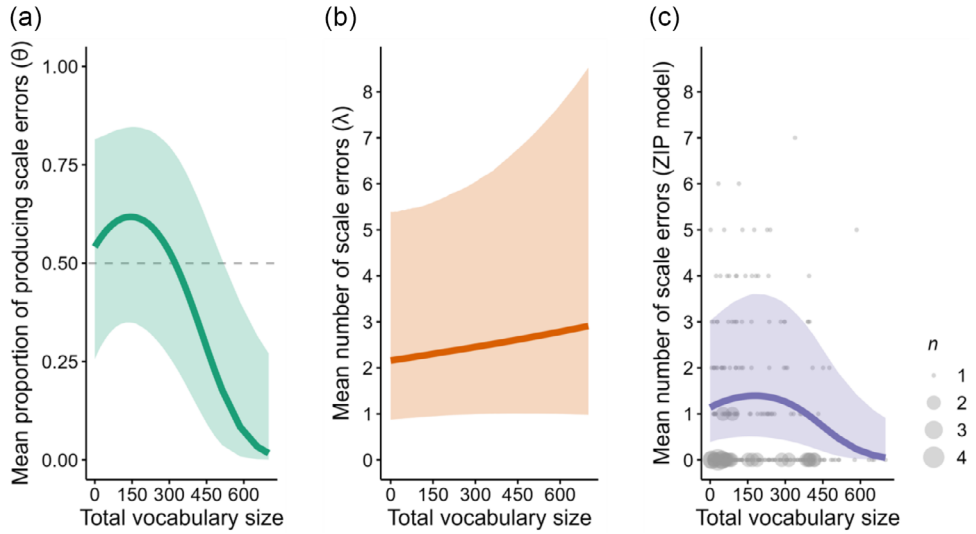

FIGURE 4 Developmental changes in scale errors drawn from the best model in Analysis 2 (total vocabulary size, Model 3). Developmental trend for the proportion part (a), count part (b), and overall zero-inflated Poisson (c). All legends are the same as those in Figure 2.

TABLE 9 Posterior median and 95% Bayesian credible intervals (CIs) of parameter estimates for the best model in Analysis 2 (total vocabulary size, Model 3).

Parameter	Posterior median [95% CI]
Logistic part of the model (θ)	
Intercept	0.44 [−0.63, 1.63]
Vocabulary size	−0.20 [−0.61, 0.20]
Vocabulary size^2	**−0.36 [−0.76, −0.02]**
Country (UK)	−0.19 [−1.95, 1.47]
Session	0.68 [−0.59, 2.98]
# Objects	−0.07 [−0.81, 0.69]
Task duration	−0.20 [−1.95, 1.49]
Gender (girl)	0.23 [−0.41, 0.86]
Count part of the model (λ)	
Intercept	0.80 [0.22, 1.37]
Age	0.07 [−0.12, 0.24]
Country (UK)	−0.17 [−1.80, 1.40]
Session	**−1.02 [−1.73, −0.33]**
# Objects	−0.12 [−0.54, 0.27]
Task duration	−0.18 [−1.78, 1.42]
Gender (girl)	0.28 [−0.03, 0.60]

Note: Children's vocabulary size was standardized. The detected effects are indicated in bold.

(MED = 0.28, CI[−0.03, 0.60]), which differed from that in Analysis 1, but detected the effect of the session (MED = −1.02, CI[−1.73, −0.33]).

The expected values of θ, λ, and y were calculated as in Analysis 1 (Figure 4). The probability of producing scale errors (i.e., θ) peaked at a vocabulary size of approximately 140 (MED = 0.62, CI[0.35, 0.85]), differing from when the vocabulary size reached the largest value (the difference in the expected value between vocabulary sizes of 140 and 700, $MED_{700-140} = 0.58$, CI[0.26, 0.81]), although the difference in the expected value straddled zero between vocabulary sizes of 140 and 0 ($MED_{140-0} = 0.07$, CI[−0.09, 0.25]). Meanwhile, the mean number of scale errors produced in the first session (i.e., λ) did not show a clear developmental trend from a vocabulary size of 0 (MED = 2.16, CI [0.88, 5.39]) to that of 700 (MED = 2.90, CI [0.98, 8.52]). For the expected value of y in which both proportion and count parts were mixed in the ZIP model, the mean number of scale errors observed peaked at a vocabulary size of approximately 180 (MED = 1.39, CI[0.51, 3.60]). The differences in the expected value between this peak and the smallest/largest vocabulary size were compatible with those in θ ($MED_{180-0} = 0.23$, CI[−0.23, 1.02]; $MED_{180-700} = 1.27$, CI[0.35, 3.35]).

A posterior predictive check confirmed that 98.9% of the observations were included in the 95% CI of the posterior predictive distribution, [0, 5], and the Spearman's rank correlation coefficient between θ and λ did not show any specific direction (MED = −0.14, CI [−0.67, 0.44]).

4.3.2 | Subvocabulary size of nouns, verbs, and adjectives as the developmental indices

A model comparison approach demonstrated that among the different categories of subvocabulary size, the subvocabulary size of adjectives was the best predictor (Model 3, WAIC = 490.82, *SE* = 31.11; Table 8), followed by that of verbs (Model 3, WAIC = 491.94, *SE* = 31.26). For both developmental indices, the selected model included their squared term for the proportion part θ and only their linear term for the count part λ. These models were superior in terms of predictability to the model with the same structure using total vocabulary size as the developmental index (WAIC = 492.28, *SE* = 31.05). The subvocabulary size of nouns showed the worst performance, suggesting that

TABLE 10 Posterior median and 95% Bayesian credible intervals (CIs) of parameter estimates for the best model in Analysis 2 (subvocabulary size, Model 3).

	Posterior median [95% CI]	
Parameter	**Verbs**	**Adjectives**
Logistic part of the model (θ)		
Intercept	0.12 [−1.51, 1.86]	0.13 [−1.53, 1.87]
Vocabulary size	−0.14 [−0.71, 0.42]	−0.24 [−0.78, 0.27]
Vocabulary size^2	−0.40 [−0.92, 0.06]	−0.41 [−0.92, 0.05]
Country (UK)	0.07 [−1.78, 1.94]	0.03 [−1.79, 1.88]
Session	0.77 [−0.55, 3.15]	0.83 [−0.55, 3.53]
# Objects	0.09 [−0.82, 1.02]	0.17 [−0.74, 1.11]
Task duration	0.06 [−1.78, 1.96]	0.03 [−1.83, 1.88]
Gender (girl)	0.01 [−0.78, 0.77]	0.03 [−0.75, 0.80]
Count part of the model (λ)		
Intercept	0.21 [−1.39, 1.90]	0.21 [−1.40, 1.91]
Age	0.12 [−0.13, 0.35]	0.13 [−0.11, 0.35]
Country (UK)	0.11 [−1.71, 1.99]	0.13 [−1.70, 2.00]
Session	**−0.98 [−1.71, −0.28]**	**−1.00 [−1.71, −0.32]**
# Objects	0.19 [−0.67, 1.02]	0.17 [−0.69, 1.00]
Task duration	0.10 [−1.68, 1.96]	0.11 [−1.69, 2.00]
Gender (girl)	0.21 [−0.17, 0.61]	0.21 [−0.18, 0.61]

Note: Children's vocabulary size was standardized. The detected effects are indicated in bold.

developmental changes in scale errors accompany the development of children's lexical abilities related to predicates rather than concrete nouns.

Although the proportion part of the selected model included the squared term of vocabulary size, its 95% CI straddled zero (adjectives, MED = −0.41, CI[−0.92, 0.05]; verbs, MED; Table 10). For the covariates, repeated exposure to the same toy sets negatively influenced the count part of scale errors (adjectives, MED = −1.00, CI[−1.71, −0.32] verbs, MED = −0.98, CI[−1.71, −0.28]). We did not detect any other clear fixed effects.

The expected values of θ, λ, and y were calculated as in the analysis of total vocabulary size. For the best model treating adjective vocabulary size as the developmental index, the peaks of θ and y appeared at a vocabulary size of 12 (MED = 0.55, CI[0.21, 0.85]) and 15 (MED = 0.76, CI[0.13, 3.79]), respectively (see Figure S2). Because the quadratic term was not included, there was no peak for λ. A clear decrease from the peak to the largest vocabulary size in the probability and the observed number of scale errors was detected (the difference in expected θ, $\text{MED}_{12-62} = 0.51$, CI[0.15, 0.82]; y, $\text{MED}_{15-62} = 0.67$, CI[0.04, 3.45]), but not a clear increase from the vocabulary size of zero to the peak point (θ, $\text{MED}_{12-0} = 0.05$, CI[−0.13, 0.82]; y, $\text{MED}_{15-0} = 0.14$, CI[−0.18, 1.00]). Compatible results were obtained for the best model, which treated verb vocabulary size as a developmental index. The expected values of θ and y peaked at vocabulary sizes of 20 (MED = 0.55, CI[0.21, 0.85]) and 30 (MED = 0.77, CI[0.14, 3.80]), respectively (Figure S3). These expected values apparently decreased after the peak toward the maximum range of verb vocabulary size (θ, $\text{MED}_{20-102} = 0.48$, CI[0.11, 0.79]; y, $\text{MED}_{30-102} = 0.63$, CI[0.01, 3.26]), but no increase toward the peak was detected (θ, $\text{MED}_{20-0} = 0.05$, CI[−0.11, 0.23]; y, $\text{MED}_{39-0} = 0.13$, CI[−0.21, 1.01]). As in the previous analysis, 97.5% of the actual data were within the 95% CI of posterior predictive distribution of [0, 4], and no correlation was detected between θ and λ.

4.4 | Discussion

For the in-lab data, we confirmed that the process during which scale errors were generated would be consistent, regardless of which age or vocabulary size was used as a developmental index. That is, an inverted U-shaped curve fits the proportion part but not the count part. In the analysis using total vocabulary size, the quadratic term was detected as negative, suggesting a clear developmental peak in the probability of scale errors. This peak appeared at a total vocabulary size of approximately 140, which was much later than the vocabulary size at which shape bias noticeably developed (Gershkoff-Stowe & Smith, 2004). Hence, if scale errors are developmentally related to the acquisition of shape bias, the causal direction is that shape bias influences the probability of scale error occurrence, creating a sensitive period of scale errors. However, the finding that the peaks between scale errors and shape bias were not accompanied makes it difficult to assume direct developmental relationships between these two because these relationships have been assumed to occur during the emergent phase of object-action association (Grzyb, Nagai, et al., 2019) or when the

lexical system rapidly changes (Grzyb, Cangelosi et al., 2019). This is also supported by the finding that the subvocabulary size of nouns had the lowest predictability compared to other parts of speech.

Instead, our model comparison approach elicited the possibility that the lexical development of predicates such as adjectives or verbs, rather than nouns, was more related to scale error occurrence. This is compatible with the findings of Hagihara et al. (2022b), who suggested the potency of verb vocabulary size as a developmental marker for scale errors. However, the quadratic terms of subvocabulary size for the proportion part straddled zero for both the best models, using adjectives or verbs as developmental indices. This failure to detect a clear developmental peak in scale errors could be because the sample size was reduced for this particular analysis from that used for the total vocabulary size from 271 to 197. In addition, data collection from children with smaller vocabulary sizes may allow for more sensitive detection of peaks.

We explored whether the vocabulary size of specific sets of words (e.g., size-related words, *big* and *little*) was related to scale errors (Supplementary Analysis 1). However, none of the specific word sets referring to the target objects, object-specific actions, or sizes was associated with scale errors. Thus, it is more likely that some general cognitive and/or linguistic mechanisms reflected in vocabulary size influence scale errors rather than specific lexical knowledge directly related to scale error tasks.

5 | GENERAL DISCUSSION

This is the first meta-analytic study using various existing datasets exploring scale errors, aiming to describe the developmental change in scale errors. Our approach revealed that the overall developmental trend of scale errors was well documented by an inverted U-shaped curve rather than a simple linear function, although nonlinearity belonged to different aspects of the scale errors between the laboratory and classroom data. At least in the in-lab data, the selected models were always the same, regardless of the developmental indices. The results of this study suggest that the nonlinear developmental change cannot be attributed to the number of scale errors itself. Rather, the probability of producing scale errors changes nonlinearly with development, and as a result, the number of observed scale errors appears to show a nonlinear trend as well. The ZIP model improved the explainability of developmental mechanisms underlying scale errors.

Another key finding of this study is that language development related to predicate vocabulary (i.e., adjectives or verbs), rather than noun vocabulary, is an important clue to understanding the developmental mechanisms of scale errors. Children's lexical development has been considered to crucially affect scale errors (Grzyb et al., 2014; Grzyb, Cangelosi, et al., 2019; Grzyb, Nagai, et al., 2019; Hagihara et al., 2022b; Hunley & Hahn, 2016; Oláh et al., 2016). Although the rapid increase in noun vocabulary size is a key marker reflecting changes in children's language abilities and most studies focused on basic abilities contributing to learning nouns efficiently (Gershkoff-Stowe & Smith, 2004; Landau et al., 1988; Smith et al., 2002), we expect that scale errors are driven by a later change in those abilities. We posit that scale errors are more likely to be produced when children develop the ability to conceptually or linguistically dissociate object-related features from objects themselves, such as semantic differentiation or abstraction (Hagihara & Sakagami, 2020; Hagihara, Yamamoto, et al., 2022; Werner & Kaplan, 1963). Acquiring language abilities, such as forming abstract concepts of object properties or actions, regardless of what objects are given may increase the probability of scale errors, as property or action concepts are sometimes activated exceedingly or too little, leading to the discarding of different aspects of objects.

Among the covariates, experiencing scale error tasks repeatedly and consistently showed a negative effect on the number of scale errors, and the gender effect was often detected for the same part of the ZIP model. Regarding the gender effect, we exploratorily assessed its robustness by treating another possible measurement of scale errors, the duration for which children engaged in producing scale errors (Supplementary Analysis 2). The results indicated that girls engaged for longer in performing scale errors than boys per session for the in-lab data, which was compatible with the results of Analysis 1; however, this effect was not observed for the classroom data.

The ZIP model is particularly helpful for investigating low-frequency data, containing a lot of zeroes. By leveraging the ZIP model, a developmental scientist can discern the factors related to the process of scale errors. Some factors may affect only the logistic or count part of the ZIP model, whereas others may be related to both parts (Atkins & Gallop, 2007). At least for the in-lab experiments, we hypothesized that developmental changes in children's cognitive/language abilities lead to a change in the probability of whether the child produces scale errors itself (i.e., the logistic part), while individual differences in children's abilities and/or task settings/procedures affect the number of scale errors (i.e., the count part). During infancy and toddlerhood, cognitive and linguistic abilities vary greatly across individuals, even among children of the same age (e.g., Frank et al., 2021). Therefore, when performing a scale error task, it is desirable to collect and analyze other indicators of children's cognitive and/or linguistic development that are thought to be relevant to this phenomenon. A possible but not fully investigated factor affecting the count part of scale errors would be children's inhibitory control. Children with immature inhibitory control are assumed to repeatedly produce scale errors, as they cannot quickly switch how to interact with miniature objects (Rosengren et al., 2010). Although parental reports are often used to assess this ability, Hagihara et al. (2022b) did not detect its effect on scale errors. Our other exploratory analysis predicting the duration of scale errors from parental reports on children's inhibitory control measures also yielded negative results (Supplementary Analysis 3). Such parent-report measures might not have sufficient sensitivity to capture individual differences in inhibitory control, and more fine-grained behavioral measures might contribute to assessing this hypothesis. We did not find any relationship between the logistic and count parts of scale errors, suggesting that these two measurements reflect different processes of scale errors. These findings emphasize the importance of scale error research in collecting several different aspects of children's abilities and analyzing them using the ZIP model.

Although we performed analyses using the ZIP model, methodological improvements are worth considering in future studies. For instance, random effects can be implemented in the ZIP model (Min & Agresti, 2005; UCLA: Statistical Consulting Group, n.d.). Although the quadratic function of the developmental indices was selected as the best model, we do not claim that this function best describes scale errors because there are various nonlinear functions and the quadratic function has many constraints, such as symmetricity. However, we believe that our approach is helpful in understanding the developmental change in scale errors because the results suggest how scale errors developmentally change can be depicted with a relatively simple nonlinear function, given the superiority of quadratic functions to more flexible functions such as B-splines (Supplementary Analysis 4), and because the negative value of the quadratic term indicates which developmental axis scale errors peak. Peak detection using quadratic functions has been used in developmental studies (e.g., Yamamoto et al., 2019).

Future research using a model comparison approach with the ZIP model would be beneficial for providing a unified account of the mechanisms underlying scale errors. Since scale errors have been considered to be related to several different developmental domains, such as inhibitory control (DeLoache et al., 2013; Ishibashi & Moriguchi, 2021; Rivière et al., 2020), size perception/comprehension (Brownell et al., 2007; Grzyb et al., 2017; Ishibashi & Moriguchi, 2017; Ware et al., 2006), and lexical development (Grzyb et al., 2014; Grzyb, Cangelosi, et al., 2019; Grzyb, Nagai, et al., 2019; Hagihara et al., 2022b; Hunley & Hahn, 2016; Oláh et al., 2016), a model comparison approach with the ZIP model will disentangle which developmental aspect relates to which part of the scale errors. To clearly detect a nonlinear developmental trend in scale errors, it is necessary to include younger children or those with smaller vocabulary sizes. Notably, considering the laboratory scale errors peaked at the age of 18–20 months in this study and children began producing scale errors around the age of 12 months (Ware et al., 2010), we suggest considering the inclusion of the 12–18 months age range. This would likely enhance the detectability of a clear increase in the probability of scale errors leading to a peak, and thus be helpful for obtaining conclusive evidence on whether scale errors developmentally change nonlinearly. Another recommendation would be to collect scale error data using different methods, such as a combination of laboratory experiments and prospective parental questionnaires from the same children. This would not only avoid underestimating the low-frequent events (Rosengren, Gutiérrez, et al., 2009; Rosengren et al., 2010) but lead to obtaining clues to explain differences in scale errors between laboratory and everyday contexts.

This meta-analytic study provides more detailed and comprehensive findings, contributing to a deeper understanding of the developmental trends of scale errors. We believe that our achievements set a fruitful stage for future research that directly investigates the relationships between scale errors and other cognitive/language abilities, leading to a unified account of the mechanisms underlying scale errors. We also speculate that the approach introduced in this study would have greater versatility for use beyond scale error research and should be encouraged in the field of developmental science. In particular, the ZIP model is expected to be a powerful analytical tool when dealing with relatively rare events that few children show, or events that are developmentally prone to zero observations. For instance, other types of action errors (Jiang & Rosengren, 2018) can be well-documented using the ZIP model. The count data of infants' word utterances in a natural setting would be another good example that can be addressed well by the model because zero observations can happen either because a child still does not speak or because a child does not talk despite already speaking. We believe that this study will be a good foundation for such studies.

CRediT Statement

Hiromichi Hagihara: Conceptualization; Methodology; Formal analysis; Investigation; Resources; Writing—original draft; Writing—review and editing; Visualization; Project administration; Funding acquisition. **Mikako Ishibashi**: Conceptualization; Methodology; Validation; Investigation; Resources; Writing—original draft; Writing—review and editing; Project administration. **Yusuke Moriguchi**: Conceptualization; Methodology; Writing—review and editing; Supervision. **Yuta Shinya**: Methodology; Resources; Writing—review and editing; Project administration; Funding acquisition.

ACKNOWLEDGMENTS

We are grateful to Katherine E. Twomey, Gert Westermann, Izumi Uehara, Beata J. Grzyb, Angelo Cangelosi, Allegra Cattani, Caroline Floccia, Martha E. Arterberry, Susan J. Hespos, Cole A. Walsh, Carolyn I. Daniels, Karl S. Rosengren, Caitlin Carmichael, Stevie S. Schein, Kathy N. Anderson, and Isabel T. Gutiérrez for sharing the datasets with us and for approving their use in this study. We also thank Keiichi Fukaya, Hiroki Yamamoto, Tenchi Mizutani, and Sho Tsuji for their invaluable comments on methodologies and manuscript organization. This research was supported by JSPS KAKENHI Grant Numbers JP18J21948, JP22KJ0525, and JP22K13664, and the Center for Early Childhood Development, Education, and Policy Research (Cedep), Graduate School of Education, The University of Tokyo.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The datasets and codes used in this study are available in the GitHub repository (https://github.com/hagi-hara/ZIP-for-scale-errors).

ORCID

Hiromichi Hagihara https://orcid.org/0000-0003-3316-600X
Yusuke Moriguchi https://orcid.org/0000-0002-9002-7834
Yuta Shinya https://orcid.org/0000-0002-5229-1554

ENDNOTES

[1] The participants in Ishibashi and Uehara (2020) partially overlapped with those in the Japanese sample in Ishibashi et al. (2021). In the present study, duplicate participants were excluded. One participant was coded

to produce scale errors in Ishibashi et al. (2021), but not in Ishibashi and Uehara (2020), due to different coding schemes and raters. We used the conservative coding in this study, that is, we treated the child as the one who did not produce any scale errors.

[2] Two additional data points in which the children's ages in months were not recorded were excluded from the final dataset.

[3] Strictly speaking, Rosengren et al. (2010) explained that the observers recorded children's actions when they possibly were scale errors. Hence, other interactions with the target objects might have been dropped from the raw data, leading to overestimating the proportion of scale error occurrences.

[4] Total vocabulary size for children who participated in Ishibashi et al. (2021) and Ishibashi and Uehara (2020) were presented at the conferences (Ishibashi & Uehara, 2017, 2019), primarily focusing on the relationships between their lexical development and pretending behavior.

[5] The J-MCDI included 281 items for concrete nouns, 103 items for verbs, and 63 items for adjectives. The O-CDI included 184 items for concrete nouns, 70 items for verbs, and 39 items for adjectives.

[6] We preliminarily confirmed if the peak at which scale errors were most likely to occur largely varied between Japan and the UK samples as the total number of items contained in the J-MCDI and O-CDI was different. As there were no large differences, we merged the two-country data to maximize the sample size while including the covariate of country, which could also reflect the different questionnaires.

REFERENCES

Arterberry, M. E., Hespos, S. J., Walsh, C. A., & Daniels, C. I. (2020). Integration of thought and action continued: Scale errors and categorization in toddlers. *Infancy*, *25*(6), 851–870. https://doi.org/10.1111/infa.12364

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology*, *21*(4), 726–735. https://doi.org/10.1037/0893-3200.21.4.726

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Bigham, S., & Bourchier-Sutton, A. (2007). The decontextualization of form and function in the development of pretence. *British Journal of Developmental Psychology*, *25*(3), 335–351. https://doi.org/10.1348/026151006X153154

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(2), 195–209. https://doi.org/10.1111/1467-985X.00130

Brownell, C. A., Zerwas, S., & Ramani, G. B. (2007). 'So big': The development of body self-awareness in toddlers. *Child Development*, *78*(5), 1426–1440. https://doi.org/10.1111/j.1467-8624.2007.01075.x

Bujang, M. A., Sa'at, N., Bakar, T. M. I. T. A., & Joo, L. C. (2018). Sample size guidelines for logistic regression from observational studies with large population: Emphasis on the accuracy between statistics and parameters based on real life clinical data. *The Malaysian Journal of Medical Sciences*, *25*(4), 122–130. https://doi.org/10.21315/mjms2018.25.4.12

Caselli, C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language*, *26*(1), 69–111. https://doi.org/10.1017/S0305000998003687

Casler, K., Eshleman, A., Greene, K., & Terziyan, T. (2011). Children's scale errors with tools. *Developmental Psychology*, *47*(3), 857–866. https://doi.org/10.1037/a0021174

DeLoache, J. S., LoBue, V., Vanderborght, M., & Chiong, C. (2013). On the validity and robustness of the scale error phenomenon in early childhood. *Infant Behavior and Development*, *36*(1), 63–70. https://doi.org/10.1016/j.infbeh.2012.10.007

DeLoache, J. S., Uttal, D. H., & Rosengren, K. S. (2004). Scale errors offer evidence for a perception-action dissociation early in life. *Science*, *304*(5673), 1027–1029. https://doi.org/10.1126/science.1093567

Elder, J. L., & Pederson, D. R. (1978). Preschool children's use of objects in symbolic play. *Child Development*, *49*(2), 500–504. https://doi.org/10.2307/1128716

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M., & Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology*, *30*(2), 326–343. https://doi.org/10.1111/j.2044-835X.2011.02042.x

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank Project*. MIT Press.

Gabry, J., & Češnovar, R. (2020). *CmdStanR: R interface to 'CmdStan'*. R package version 0.3.0. Retrieved from https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Develop ment*, *75*(4), 1098–1114. https://doi.org/10.1111/j.1467-8624.2004.00728.x

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Grzyb, B. J., Cangelosi, A., Cattani, A., & Floccia, C. (2017). Decreased attention to object size information in scale errors performers. *Infant Behavior and Development*, *47*, 72–82. https://doi.org/10.1016/j.infbeh.2017.03.001

Grzyb, B. J., Cangelosi, A., Cattani, A., & Floccia, C. (2019). Children's scale errors: A by-product of lexical development? *Developmental Science*, *22*(2), e12741. https://doi.org/10.1111/desc.12741

Grzyb, B. J., Cattani, A., Cangelosi, A., & Floccia, C. (2014). Children in a wonderland: How language and scale errors may be linked. In *4th International Conference on Development and Learning and on Epigenetic Robotics* (pp. 269–274). https://doi.org/10.1109/DEVLRN.2014.6982992

Grzyb, B. J., Nagai, Y., Asada, M., Cattani, A., Floccia, C., & Cangelosi, A. (2019). Children's scale errors are a natural consequence of learning to associate objects with actions: A computational model. *Developmental Science*, *22*(4), e12777. https://doi.org/10.1111/desc.12777

Hagihara, H., Ishibashi, M., Moriguchi, Y., & Shinya, Y. (2022a). Data from "Object labeling activates young children's scale errors at an early stage of verb vocabulary growth". *Journal of Open Psychology Data*, *10*, 15. https://doi.org/10.5334/jopd.70

Hagihara, H., Ishibashi, M., Moriguchi, Y., & Shinya, Y. (2022b). Object labeling activates young children's scale errors at an early stage of verb vocabulary growth. *Journal of Experimental Child Psychology*, *222*, 105471. https://doi.org/10.1016/j.jecp.2022.105471

Hagihara, H., & Sakagami, M. (2020). Initial noun meanings do not differentiate into object categories: An experimental approach to Werner and Kaplan's hypothesis. *Journal of Experimental Child Psychology*, *190*, 104710. https://doi.org/10.1016/j.jecp.2019.104710

Hagihara, H., Yamamoto, H., Moriguchi, Y., & Sakagami, M. (2022). When "shoe" becomes free from "putting on": The link between early meanings of object words and object-specific actions. *Cognition*, *226*, 105177. https://doi.org/10.1016/j.cognition.2022.105177

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, *27*(3), 689–705. https://doi.org/10.1017/S0305000900004414

Hu, M. C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse*, *37*(5), 367–375. https://doi.org/10.3109/00952990.2011.597280

Hunley, S. B., & Hahn, E. R. (2016). Labels affect preschoolers' tool-based scale errors. *Journal of Experimental Child Psychology*, *151*, 40–50. https://doi.org/10.1016/j.jecp.2016.01.007

Ishibashi, M., & Moriguchi, Y. (2017). Understanding why children commit scale errors: Scale error and its relation to action planning and inhibitory control, and the concept of size. *Frontiers in Psychology*, *8*, 826. https://doi.org/10.3389/fpsyg.2017.00826

Ishibashi, M., & Moriguchi, Y. (2021). Neural basis of scale errors in young children. *Developmental Neuropsychology*, *46*(2), 109–120. https://doi.org/10.1080/87565641.2021.1887871

Ishibashi, M., Twomey, K. E., Westermann, G., & Uehara, I. (2021). Children's scale errors and object processing: Early evidence for cross-cultural differences. *Infant Behavior and Development*, *65*, 101631. https://doi.org/10.1016/j.infbeh.2021.101631

Ishibashi, M., & Uehara, I. (2017). Children's scale errors: Its relationship to semantic knowledge and pretending behaviors. Poster presented at the Lancaster International Conference on Infant and Early Child Development, Lancaster, UK.

Ishibashi, M., & Uehara, I. (2019) Children's scale errors: Developmental changes in pretending and language comprehension. Poster presented at the 19th annual meeting of the Japanese Society of Baby Science, Tokyo, Japan (in Japanese).

Ishibashi, M., & Uehara, I. (2020). The relationship between children's scale error production and play patterns including pretend play. *Frontiers in Psychology*, *11*, 1776. https://doi.org/10.3389/fpsyg.2020.01776

Jiang, M. J., & Rosengren, K. S. (2018). Action errors: A window into the early development of perception–action system. *Advances in Child Development and Behavior*, *55*, 145–171. https://doi.org/10.1016/bs.acdb.2018.04.002

Jones, S. S. (2003). Late talkers show no shape bias in a novel name extension task. *Developmental Science*, *6*(5), 477–483. https://doi.org/10.1111/1467-7687.00304

Karazsia, B. T., & van Dulmen, M. H. (2008). Regression models for count data: Illustrations using longitudinal predictors of childhood injury. *Journal of Pediatric Psychology*, *33*(10), 1076–1084. https://doi.org/10.1093/jpepsy/jsn055

Kemler Nelson, D. G., Russell, R., Duke, N., & Jones, K. (2000). Two-year-olds will name artifacts by their functions. *Child Development*, *71*(5), 1271–1288. https://doi.org/10.1111/1467-8624.00228

Kobayashi, H. (1997). The role of actions in making inferences about the shape and material of solid objects among Japanese 2 year-old children. *Cognition*, *63*(3), 251–269. https://doi.org/10.1016/S0010-0277(97)00007-3

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14. https://doi.org/10.1080/00401706.1992.10485228

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191. https://doi.org/10.1016/j.cognition.2020.104191

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, *65*(1), 163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*(11), 1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Chapman and Hall/CRC.

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, *5*(1), 1–19. https://doi.org/10.1191/1471082X05st084oa

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., & Altman, D. G. (2011). Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, *32*(3), 437–440. https://doi.org/10.3174/ajnr.A2425

Ogura, T., & Watamaki, T. (2004). *Nihongo MacArthur Nyuyoji Gengo Hattatsu Shitumonshi [Japanese MacArthur communicative development inventories]*. Kyoto International Social Welfare Exchange Centre.

Ogura, T., Watamaki, T., & Inaba, T. (2016*). Nihongo MacArthur nyuyoji gengo hattatsu shitumonshi no kaihatsu to kenkyu [The development and research of Japanese MacArthur communicative development inventories]*. Nakanishiya Shuppan.

Oláh, K., Elekes, F., Pető, R., Peres, K., & Király, I. (2016). 3-year-old children selectively generalize object functions following a demonstration from a linguistic in-group member: Evidence from the phenomenon of scale error. *Frontiers in Psychology*, *7*, 963. https://doi.org/10.3389/fpsyg.2016.00963

Rivière, J., Brisson, J., & Aubertin, E. (2020). The interaction between impulsivity, inhibitory control and scale errors in toddlers. *European Journal of Developmental Psychology*, *17*(2), 231–245. https://doi.org/10.1080/17405629.2019.1567324

Rosengren, K. S. (n.d.). Variability in young children's interactions with scale replicas: Exploratory play, general play, pretense, and scale errors. Unpublished data.

Rosengren, K. S., Carmichael, C., Schein, S. S., Anderson, K. N., & Gutiérrez, I. T. (2009). A method for eliciting scale errors in preschool classrooms. *Infant Behavior and Development*, *32*(3), 286–290. https://doi.org/10.1016/j.infbeh.2009.03.001

Rosengren, K. S., Gutiérrez, I. T., Anderson, K. N., & Schein, S. S. (2009). Parental reports of children's scale errors in everyday life. *Child Development*, *80*(6), 1586–1591. https://doi.org/10.1111/j.1467-8624.2009.01355.x

Rosengren, K. S., Schein, S. S., & Gutiérrez, I. T. (2010). Individual differences in children's production of scale errors. *Infant Behavior and Development*, *33*(3), 309–313. https://doi.org/10.1016/j.infbeh.2010.03.011

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*(1), 127–141. https://doi.org/10.1002/sim.2331

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, *25*(4), 666–678. https://doi.org/10.1016/j.jcps.2015.04.004

Saji, N., Imai, M., Saalbach, H., Zhang, Y., Shu, H., & Okada, H. (2011). Word learning does not end at fast-mapping: Evolution of verb meanings through reorganization of an entire semantic domain. *Cognition*, *118*(1), 45–61. https://doi.org/10.1016/j.cognition.2010.09.007

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19. https://doi.org/10.1111/1467-9280.00403

Stan Development Team. (2021). *Stan modeling language users guide and reference manual*. version 2.28.2. Retrieved from https://mc-stan.org

UCLA: Statistical Consulting Group. (n.d.). *How do I run a random effect zero-inflated Poisson model using nlmixed?* https://stats.oarc.ucla.edu/sas/faq/how-do-i-run-a-random-effect-zero-inflated-poisson-model-using-nlmixed

Ware, E. A., Uttal, D. H., & DeLoache, J. S. (2010). Everyday scale errors. *Developmental Science*, *13*(1), 28–36. https://doi.org/10.1111/j.1467-7687.2009.00853.x

Ware, E. A., Uttal, D. H., Wetter, E. K., & DeLoache, J. S. (2006). Young children make scale errors when playing with dolls. *Developmental Science*, *9*(1), 40–45. https://doi.org/10.1111/j.1467-7687.2005.00461.x

Watanabe, S. (2010a). Equations of states in singular statistical estimation. *Neural Networks*, *23*(1), 20–34. https://doi.org/10.1016/j.neunet.2009.08.002

Watanabe, S. (2010b). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594. http://jmlr.org/papers/v11/watanabe10a.html

Werner, H., & Kaplan, B. (1963). *Symbol formation: An organismic-developmental approach to language and the expression of thought*. John Wiley.

Wiesner, M., & Kim, H. K. (2006). Co-occurring delinquency and depressive symptoms of adolescent boys and girls: A dual trajectory modeling approach. *Developmental Psychology*, *42*(6), 1220–1235. https://doi.org/10.1037/0012-1649.42.6.1220

Yamamoto, H., Sato, A., & Itakura, S. (2019). Eye tracking in an everyday environment reveals the interpersonal distance that affords infant-parent gaze communication. *Scientific Reports*, *9*, 10352. https://doi.org/10.1038/s41598-019-46650-6

Zuniga-Montanez, C., Kita, S., Aussems, S., & Krott, A. (2021). Beyond the shape of things: Infants can be taught to generalize nouns by objects' functions. *Psychological Science*, *32*(7), 1073–1085. https://doi.org/10.1177/0956797621993107

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hagihara, H., Ishibashi, M., Moriguchi, Y., & Shinya, Y. (2024). Large-scale data decipher children's scale errors: A meta-analytic approach using the Zero-Inflated Poisson models. *Developmental Science*, e13499. https://doi.org/10.1111/desc.13499