Probabilistic Modeling of Variation in Pilot Performance During Flight Training

Kento Yamada, Harumi Ikeshita, Yuta Kyoya, and Makoto Ueno

***Abstract*—A probabilistic analysis for the variability of pilot performance was performed to model the variation in pilot performance during flight training from the viewpoint of reliability theory. We summarized flags among all applicants tallied in flight training to a histogram. Various probability distributions were fitted to the histogram using two bootstrap goodness-of-fit tests. We found that a limit of the marginal distribution of Ryu's bivariate exponential distribution gave the best approximation of the histogram. Defining a random variable for the conditional hazard function as the training step was the key interpreting the physical background of the fitted distribution in terms of the growth process during training. Its hazard function showed keeping the number of flags per flight within a few was important. Also, calculating the ratio to the expectation for each training step and visualizing the transition of the cumulative number of flags revealed a concave growth model as the basic process lying in the background. Moreover, fundamental assumptions of software reliability growth model (SRGM) were interpreted in terms of pilot training, and the existence of a stochastic process was discussed. Visualizing personal processes appearing in reality, we found that their shapes were similar to those of SRGM. Therefore, applying SRGM to pilot training data is expected in the future.**

***Index Terms*—Competency-based training, flight training, performance analysis, pilot performance, reliability, resilience, statistical analysis, stochastic processes.**

NOMENCLATURE

AM	Assessment marker.
BVE	Bivariate exponential distribution.
CBCT	Competency-based check and training.
CBT	Competency-based training.
CDF	Cumulative distribution function.
GoF	Goodness-of-fit.
ICAO	International Civil Aviation Organization.
JAL	Japan Airlines, Co., Ltd.
LMR	Limit of the marginal distribution of Ryu's BVE.
PDF	Probability density function.
PMF	Probability mass function.
QBT	Quantity-based training.
SRGM	Software reliability growth model.

Notation

$\mathrm{Cdf}_{\mathrm{emp}}$	Empirical CDF.
$\mathrm{Cdf}_{\mathrm{fit}}$	Fitted CDF.
$\mathrm{Cdf}_{\mathrm{LMR}}$	CDF for the LMR.
$\mathrm{Cdf}^{-1}_{\mathrm{LMR}}$	Inverse function of CDF for the LMR.
$\mathrm{Cdf}_{\mathrm{Ryu}}$	CDF for the marginal distribution of the Ryu's BVE.
$\mathrm{E}_{\mathrm{LMR}}$	Expectation for the LMR.
$\mathrm{h}_{\mathrm{LMR}}$	Hazard function for the LMR.
$\mathrm{pdf}_{\mathrm{cg}}$	PDF for the compound gamma distribution.
$\mathrm{pdf}_{\mathrm{gam}}$	PDF for the gamma distribution.
$\mathrm{pdf}_{\mathrm{LMR}}$	PDF for the LMR.
$\mathrm{pmf}_{\mathrm{fit}}$	Fitted PMF.
$\mathrm{Sf}_{\mathrm{LMR}}$	Survival function for the LMR.
a	Number of flights required to progress the training step.
d_{χ^2}	Test statistic of chi-squared test.
d_{KS}	Test statistic of Kolmogorov–Smirnov test.
i	Nonnegative integer.
k	Step of training progress.
K	Function that returns k for N.
M	Mean growth model of t for N.
n	Number of performance flags per flight.
n_{cum}	Cumulative number of performance flags.
N	Number of flights.
N_{est}	Estimated number of flights.
N_{obs}	Observed number of flights.
N_{total}	Total number of flights.
t	Continuous variable of n.
W	Lambert W function.
α, β, q	Parameters of the compound gamma distribution.
γ	Number of groups for chi-squared test.
Γ	Gamma function.
Γ_i	Incomplete gamma function.
λ', λ, s	Parameters of the marginal distribution of Ryu's BVE.

Manuscript received 29 June 2022; revised 19 September 2022 and 9 February 2023; accepted 5 July 2023. This work was supported by JSPS KAKENHI under Grant JP21K14350. Associate Editor: T. Dohi. *(Corresponding author: Kento Yamada.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Research Institute of Human Engineering for Quality Life under Application No. E20-28-1.

Kento Yamada and Makoto Ueno are with the Japan Aerospace Exploration Agency, Tokyo 182-8522, Japan (e-mail: yamada.kento@jaxa.jp; ueno.makoto@jaxa.jp).

Harumi Ikeshita and Yuta Kyoya are with the Japan Airlines Company, Ltd., Tokyo 144-0041, Japan (e-mail: ikeshita.q8ay@jal.com; kyoya.n3yk@jal.com).

Digital Object Identifier 10.1109/TR.2023.3294021

I. INTRODUCTION

AIRLINE safety initiatives are moving from reactive measures, such as analyzing accidents and incidents and formulating recurrence prevention measures, to proactive measures, such as analyzing pilot skills and improving their

resilience at preventing events from developing into incidents and accidents. Improving resilience requires pilot training to reflect actual aircraft operations more. In addition to the technical skills related to aircraft maneuvering, it must also address the nontechnical skills necessary for crews to cooperate and carry out safe operations. This background has led to CBT becoming mainstream worldwide [1], [2], [3].

CBT is based on the idea of giving pilots through training the required competencies as crews from the viewpoint of preventing safety problems. ICAO defines a competency as a dimension of human performance that is used to reliably predict successful performance on the job [4]. While QBT manages pilot abilities by the amount of the training administered, CBT can flexibly respond to each pilot's acquisition of competencies by clearly setting out the competencies required in actual flight operations.

However, since an actual flight operation is a series of processes that deal with various events, there are variations in the competencies exerted from day-to-day depending on uncertain factors, such as the situations faced, how to handle them appropriately, and others. Then, it is quite difficult to quantitatively measure the status of competency acquisition. Events encountered during actual flight operations are probabilistic entities that do not always occur. Although training is performed so that events are appropriately handled whenever they occur, the appropriateness of pilot handling when they occur is also stochastic. Currently, this decision is dependent on instructors.

Lots of research about pilot training are going on. They are mainly conducted in the area of human factors, and their results support that the aircraft operation includes various variabilities. Previous studies have shown that a sense-making activity is essential for properly processing events [5], [6], [7], but this becomes difficult when the pilot is in a state of surprise [8], [9] or fatigue [10]. Furthermore, it has been shown that the complexity of cockpit displays probabilistically induces pilot visual recognition errors and correspondence errors, and the autopilot is set into an inappropriate mode as a result [11]. There is also a problem with creating training that faithfully simulates actual flight operations. In conventional simulator training, it is known beforehand that a particular event will occur. The fact that actual flight operations have events that are difficult to predict also contributes to stochastic variations in pilot performance. Previous studies on this aspect suggest that it will be adequate to make simulator training scenarios unpredictable [12], [13]. Therefore, CBT faces the challenge that competencies are easily affected by stochastic processes and that it is not easy to quantify competency acquisition status.

From this background, it is crucial and practically required to develop a method to measure the competency acquisition status, including the probabilistic entities in a flight training program, but as far as we know, there is no precedent for such research in pilot training.

On the other hand, studies on stochastic human performance modeling have been conducted in reliability theory [14], [15], [16]. Many tools for the reliability analysis have been developed and applied to human performances in various practical fields, such as railway systems [17], healthcare systems, such as home-based rehabilitation [18], and so on.

The SRGM is useful for modeling the converging growth process [19]. The SRGM is a model used in software development that predicts the number of errors over the development period. It is applied to determine software release timing by estimating how many errors have been eliminated compared with to the total number of errors estimated from the past development. The expression of the model gets wider. As it starts from a simple concave shape [20], other possible shapes are constructed by introducing another parameter for a real effect, such as fault recovery efficiency [21].

Based on the previous discussion, the present research aims to construct a model that can explain and measure the individual growth process during a JAL's captain upgrade training. However, to the best of authors' knowledge, it is not even known what trajectory a growth model of an applicant in pilot training follows.

As the first step in this research, the article's contribution is that applying a fundamental analysis in reliability theory to pilot training data revealed a variation in pilot performance in standard progress, which instructors most commonly observed during the training. As a result, the following four points were found:

1) A range in which the expected numbers of observed indicators per training should be kept;
2) An ideal growth model calculated from the fitted distribution to the variation histogram assuming a simple growth process;
3) Differences in growth processes between the ideal model and individual processes appearing in real;
4) Underlying assumptions and similarity of shapes between personal growth processes and SRGM.

The rest of this article is organized as follows. Section II provides the methods of training and statistical analysis. Section III shows the result of distribution fitting. Section IV discusses the stochastic property and process in flight training. Finally, Section V concludes this article.

II. Methods

A. Participants

The participants in this research were 76 first officer applicants in a captain promotion training program and 142 captains serving as instructors. The first officers had about 5000 h of flight experience and had Commercial Pilot Licenses. In comparison, the captains had about 10 000 h of flight experience and were trained as instructors by the company. In this study, the data used for analysis consisted only of cases in which informed consent were obtained from both the first officer candidate and the captain. The data acquisition corresponds to an experiment.

B. Apparatus

Boeing 737, 767, 777, and 787 aircrafts were used in training, which was conducted on the line (that is, during actual revenue flights). The routes selected for the training covered domestic and short-haul international flights and were carefully chosen to remove bias from applicant flight experience.

TABLE I
COMPONENTS OF ASSESSMENT IN THE CBCT

Competency	AM	Description
Attitude (AA)	AAC	Exercizing cost awareness.
	ACA	Executing an operation carefully and accurately.
	ACS	Considering customer satisfaction.
	AMC	Leading teamwork for the cooperation of other crewmembers.
	ARD	Understanding and cooperating with the operations of other sections.
	ARM	Utilizing resources.
	ASA	Operating a flight safely and in compliance with laws and regulations.
	ATE	Recognizing and dealing with threats, errors, and UAS.
Application of Procedures (AP)	PAC	Carrying out ATC Communication in compliance with procedures
	PAP	Compliance with the laws, regulations, procedures, and recommendations required by the situation.
	PCC	Carrying out mutual confirmation.
	PCL	Performing Checklists.
	POE	Executing procedures understanding their objectives and essence.
	PSC	Performing Callouts as stipulated by the SOP.
Communication (CO)	C2W	Performing two-way communication.
	CAS	Insisting on safety.
	CBR	Sharing schedule and awareness.
	CIN	Questioning for safety.
	CNC	Accurately utilizing and interpreting non-verbal communication.
Decision Making (DM)	DMS	To use precise and timely processes of Decision Making.
	DOC	Making a decision choosing a solution.
	DOG	Giving all options of solutions.
	DPD	Identifying the problem.
	DRA	Assessing risk.
	DRM	Making decisions, reviewing them, and changing them as needed.
Flight Path Management, automation (FA)	FAE	Keeping the combination of speed, altitude, and position within reasonable limits.
	FAF	Making the appropriate FMS inputs and selecting the appropriate automation mode for the Phase of Flight.
	FAM	Monitoring the automation mode and flight path to achieve the desired state.
	FAW	Recognizing and avoiding cloud areas by effective use of Weather Radar and visual inspection.
Flight Path Management, manual control (FM)	FMA	Properly controlling the altitude of the aircraft to obtain desired specifications, flight paths, and tracks.
	FME	Maintaining the combination of speed, altitude, and position within reasonable limits.
	FMG	Selecting and utilizing the appropriate Flight Guidance Systems Mode that supports the phase of flight.
	FMP	Premeditatedly controlling aircraft attitude and thrust.
	FMT	Properly controlling thrust according to the situation.
	FMW	Recognizing and effectively avoiding cloud areas by effective use of Weather Radar and visual inspection.
Knowledge (KK)	KAE	Having and using knowledge of airports, terrain, and routes.
	KAP	Having and using knowledge of aircraft operation and performance.
	KAT	Having and using knowledge of ATC.
	KLM	Having and using knowledge of aviation regulations, company regulations, procedures, recommendations, etc.
	KNT	Having and using knowledge of TEM, MCC, Non-Technical Skills, and Human Performance, etc.
	KWX	Having and using knowledge of the weather necessary for operation.
Situational Awareness (SA)	SAA	Recognition of aircraft and internal conditions.
	SAL	Analyzing the situation.
	SAT	Predicting the situation.
	SEA	Recognizing the external situation of the aircraft and time.
	SFA	Recognizing the status of flight oneself and other flight crews.
Team Building (TB)	TCO	Cooperation.
	TCR	Constructively resolving conflicts.
	TLS	Showing leadership.
Workload Management (WM)	WDI	Allocating tasks.
	WPR	Prioritizing.
	WTM	Using time efficiently.

C. Count Data of Pilot Performance

The training program used in this research, CBCT, is a CBT program developed by JAL. The contents of captain upgrade training are mainly separated into the simulator and line operation training. This article analyzed the data obtained by the latter training.

The line operation training is conducted in actual operations with passengers on board. An applicant and an instructor are in the cockpit. In training, the applicant act as a captain, and the instructor acts as a first officer. The main difference between the captain and the first officer is decision-making authority. Then, simultaneously fulfilling the role of the first officer, the instructor measures the exertion of the applicant's competencies as the captain by transferring the decision-making authority to the applicant. If the training is interrupted, the decision-making authority reverts to the instructor qualified as the captain.

The exertion of the applicant's competencies is measured based on the applicant's behavior observed during training. The nature of the behavior is set to be broken down and analyzed by ten competencies. Each competency was further decomposed into AMs, which consisted of specific performances to deal with an event during the flight, as given in Table I. Every AM has a three-letter name, and its content is specifically determined. The AMs are used to measure whether or not pilots precisely exerted the competencies, and consist of a basis to treat the behavior as a vector.

Each training flight had eight phases: preflight, takeoff, climb, cruise, descent, approach, landing, and postflight. In each phase, pilot competencies were checked by an instructor based on the AMs. Flags are tallied when the instructor feels differences between the actual performance of an applicant and the expected performance of a captain with ideal competencies. The measured competencies varied by flight phase, as given in Table II. Competencies measured in the phase are symbolized by "o" while ones not measured in the phase are symbolized by "x." Since the events encountered during each flight were stochastic, pilots sometimes had no opportunity to exert a competency, and in such cases, no tally was recorded.

This article focuses on the number of flags, n, recorded in each training flight since the number of performance flags is considered one of the simplest measures between ideal and observed competencies, and it will ideally converge to zero as the training progresses. The number of flags for every competency

TABLE II
PHASES AND COMPETENCIES IN THE CBCT

Competency	Preflight	Takeoff	Climb	Cruise	Descent	Approach	Landing	After landing post flight
AA	o	o	o	o	o	o	o	o
AP	o	o	o	o	o	o	o	o
CO	o	x	o	o	o	o	x	o
DM	o	x	o	o	o	o	x	o
FA	x	o	o	o	o	o	x	x
FM	o	o	o	x	o	o	o	o
KK	o	o	o	o	o	o	o	o
SA	o	o	o	o	o	o	o	o
TB	o	x	o	o	o	o	x	o
WM	o	x	o	o	o	o	x	o

is also important, but it is summarized as the number of flags per flight to capture the macroscopic nature of the growth process. Not only is the flag concerned during training. Then, the process focused in this article is a process partially visualized via the number of flags.

Since, the count data for each applicant is not enough to fit a probability distribution to a histogram, the random variable n for an applicant in standard progress was considered instead. The count data we statistically analyzed was made by gathering the n from all the training flights flown by all applicants. These data included 6697 training flights over 35 months. Flights in which n was zero included those in which no event required the exertion of competencies occurred, and we removed such cases from our analysis. Therefore, the remaining 6353 flights were used for the analysis. A flight with $n = 0$ means that an applicant encountered at least one event to deal with during the flight, and the applicant appropriately responded to them.

The count data is essentially a discrete variable since the flag is a Boolean variable so that the instructor feels the difference or not. Considering the data as a continuous variable makes it possible to apply it to the measurement of the same flight with different indicators, thus increasing its versatility when the number of AMs changes due to the update of the training method. To get correspondence between the discrete and continuous variables, we can define the discrete variable as the range of the continuous variable. In this article, we simply define n as

$$n = \lfloor t \rfloor \tag{1}$$

where $\lfloor \cdot \rfloor$ is a floor function by which an integer n is returned from a real number in $[n, n+1)$ and t is a nonnegative real number as the continuous variable of the flag. On the contrary, when determining t from n, it is simply calculated by

$$t \approx n + \frac{1}{2} \tag{2}$$

since it is not known how t is distributed in $[n, n+1)$.

D. Preliminary Analysis

An analysis of heterogeneity was performed prior to fitting. While it is natural for histograms of n to vary according to individual competencies, the possibility exists that the histograms may vary unnaturally due to a lack of training or other reasons. Since this article aims to examine the expected variation of the applicants' competencies, we investigated the variance of the personal relative frequency of n. Then, the Grubbs' test for outliers was performed to detect those with unnaturally large variations. Fig. 1 shows the histogram of the variance of the relative frequency among all applicants. Consequently, the training data of six applicants, whose variances of the relative frequencies were large, were omitted.

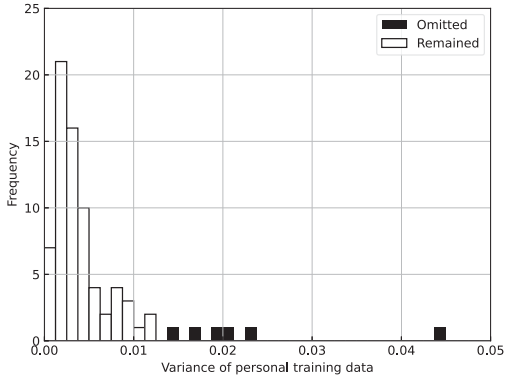

Fig. 1. Histogram of variance of the number of flags tested by Grubbs' test for outliers.

E. Distribution Fitting

We performed distribution fitting to the count data using several continuous distributions defined on the nonnegative real number and several discrete distributions defined on the nonnegative integer to determine which gave the best fit. Among the statistical functions available in SciPy [22], distributions applied to lifetime analysis were selected as candidates. As some candidate distributions are not suited to the maximum likelihood method and the experimental data are limited, we used the least-squares method of the CDF with the Powell method to minimize the sum of squared residuals. The abscissa was resolved in 0.1 increments to fit the continuous distribution to the $\mathrm{Cdf}_{\mathrm{emp}}(t)$, while the discrete distribution was fitted to the $\mathrm{Cdf}_{\mathrm{emp}}(n)$.

F. Bootstrap Goodness-of-Fit Tests

We tested the statistical data by two bootstrap GoF tests. One is a bootstrap chi-squared test, and the other is a bootstrap Kolmogorov–Smirnov test [23]. Instead of the theoretical distributions to see the significance level in the conventional

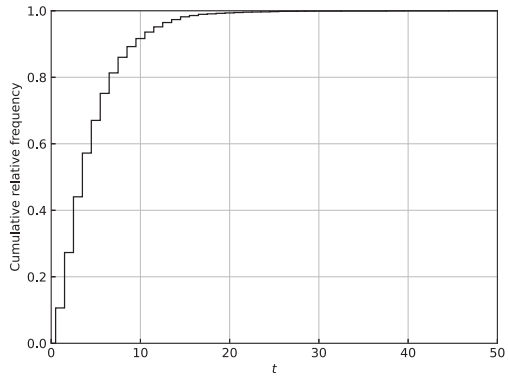

Fig. 2. $\mathrm{Cdf}_{\mathrm{emp}}(t)$.

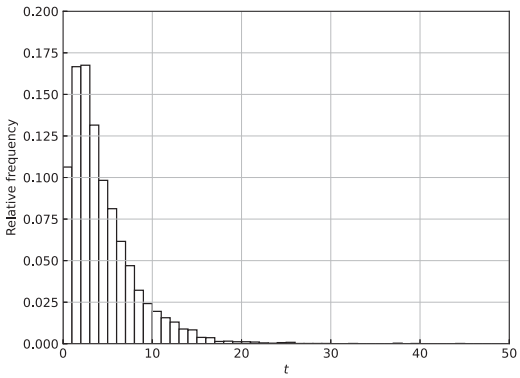

Fig. 3. Relative frequency histogram of t.

tests, the bootstrap GoF test iteratively constructs the probability distribution of test statistics by randomly resampling the same amount of data from the fitted distribution and calculating the statistics again. Testing a continuous probability distribution, we transform the resampling data to the count data using (1). The number of iterations in this article was set at 10 000. The significance level was set at 0.1.

The test statistics are d_{χ^2} and Kolmogorov–Smirnov distance, d_{KS}. We calculated d_{χ^2} as

$$d_{\chi^2} = \sum_{i=0}^{\gamma} \frac{(N_{\mathrm{obs}}(i) - N_{\mathrm{est}}(i))^2}{N_{\mathrm{est}}(i)} \tag{3}$$

where $N_{\mathrm{est}}(i) = N_{\mathrm{total}} \cdot (\mathrm{Cdf}_{\mathrm{fit}}(i+1) - \mathrm{Cdf}_{\mathrm{fit}}(i))$ and $N_{\mathrm{est}}(i) = N_{\mathrm{total}} \cdot \mathrm{pmf}_{\mathrm{fit}}(i)$ for continuous and discrete distributions, respectively. We set $\gamma = 15$ to keep $N_{\mathrm{est}}(\gamma) > 10$ [24]. As d_{KS} is calculated for the distribution of trees in a woodland [25], we calculated d_{KS} as

$$d_{\mathrm{KS}} = \max(|\mathrm{Cdf}_{\mathrm{emp}}(i) - \mathrm{Cdf}_{\mathrm{fit}}(i)|). \tag{4}$$

III. Results

A. Fitting Result

The $\mathrm{Cdf}_{\mathrm{emp}}(t)$ and relative frequency histogram of t are shown in Figs. 2 and 3, respectively. The p-values and test statistics for each distribution are given in Table III. The tests were not applied to some distributions since random variables can take very large values, making it impossible to allocate arrays or making computation time very long during resamplings in the bootstrap method. Distributions exceeding the significance level were written in bold. The probability distribution that gave the best fit was LMR, which is a limit of the marginal distribution of Ryu's BVE [26]. The fitted results of $\mathrm{Cdf}_{\mathrm{emp}}(t)$ and relative frequency histogram are shown in Figs. 4 and 5, respectively.

TABLE III
P-Values and Test Statistics by Bootstrap GoF Tests

Distribution	p_{χ^2}	d_{χ^2}	p_{KS}	d_{KS}
Betaprime	0.0	730.1	0.0	0.0630
Compound gamma	**0.1081**	**23.7**	**0.1534**	**0.0138**
Burr Type III	0.0	92.4	0.0	0.0248
Burr Type XII	0.0044	33.0	0.0355	0.0148
Chi	0.0	8488.3	0.0	0.1035
Exponential power	0.0	355.4	0.0	0.0612
Birnbaum-Saunders	0.0	174.1	0.0	0.0352
Fold Cauchy	N/A	2130.9	N/A	0.131
Gamma	0.0015	36.4	0.003	0.017
Generalized gamma	0.0164	29.0	0.006	0.0167
LMR	**0.3886**	**16.2**	**0.2593**	**0.0122**
Generalized pareto	0.0	435.6	0.0	0.0945
Gompertz	0.0	252.5	0.0	0.0593
Half cauchy	N/A	1553.4	N/A	0.180
Half logistic	0.0	170.6	0.0	0.0529
Inverse Gauss	0.0	509.1	0.0	0.0580
Lognormal	0.0	179.1	0.0	0.0351
Nakagami	0.0	238.1	0.0	0.0483
Reciprocal inverse Gauss	0.0001	61.7	0.0024	0.0180
Rice	0.0	8489.6	0.0	0.104
Weibull	0.0	81.2	0.0	0.0303
Negative binomial	0.015	58.0	0.01	0.0164
Zero modified Poisson	0.0	9445.4	0.0	0.104

The bold entities are important to show the distributions satisfying the criteria for p>0.1.

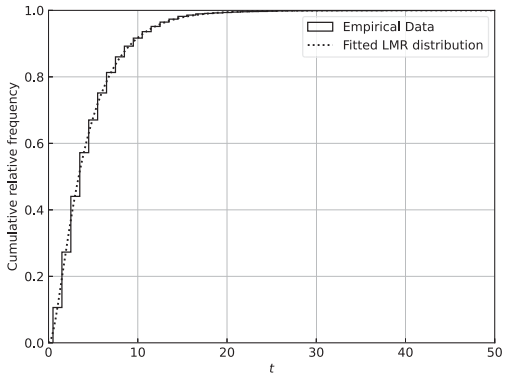

Fig. 4. $\mathrm{Cdf}_{\mathrm{emp}}(t)$ with the fitted LMR.

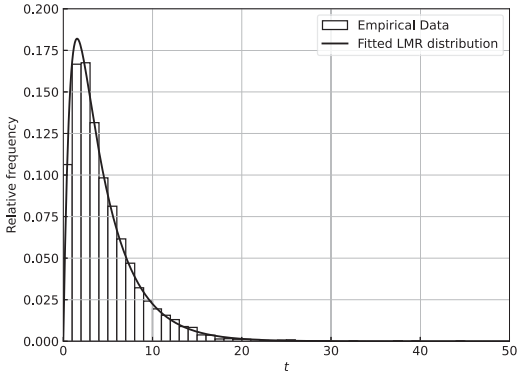

Fig. 5. Relative frequency histogram of t with the fitted LMR.

B. Limit of the Marginal Distribution of Ryu's BVE

We mention why this limit is required instead of the original distribution. The CDF of the marginal distribution of Ryu's BVE is defined as

$$\mathrm{Cdf}_{\mathrm{Ryu}}(t) = 1 - e^{-\lambda' t - \lambda t + \frac{\lambda}{s}(1-e^{-st})} \tag{5}$$

where λ', λ, and s are positive real parameters. The estimated parameters fitted to the CDF of the present data were

$$(\lambda', \lambda, s) = \left(3.22 \times 10^{-28}, 2.73 \times 10^{-1}, 1.17\right) \tag{6}$$

and the square of residuals was 1.75×10^{-6}. It suggested that $\lambda' \to 0$ but $\lambda' > 0$ from the definition. Then, we needed to consider the LMR.

The CDF and PDF of the probability distribution for $\lambda' \to 0$ are defined on the nonnegative real number and written as

$$\mathrm{Cdf}_{\mathrm{LMR}}(t) = 1 - e^{-\lambda t + \frac{\lambda}{s}(1-e^{-st})} \tag{7a}$$

$$\mathrm{pdf}_{\mathrm{LMR}}(t) = \lambda \left(1 - e^{-st}\right) e^{-\lambda t + \frac{\lambda}{s}(1-e^{-st})}. \tag{7b}$$

His paper did not discuss this limit of the probability distribution, but its survival function, $\mathrm{Sf}_{\mathrm{LMR}}(t) \equiv 1 - \mathrm{Cdf}_{\mathrm{LMR}}(t)$, was derived as an intermediate product, and the proof of the derivation was also given in the first proposition of Appendix A. The estimated parameters of this distribution to the CDF of the present data were

$$(\lambda, s) = \left(2.73 \times 10^{-1}, 1.17\right) \tag{8}$$

and the squared residual was 1.75×10^{-6}. From above, we confirmed that we obtained the same fitting result as the fitting result of the marginal distribution of the Ryu's original distribution.

We used the inverse transform method to generate random variables following this probability function from uniformly distributed random variables. As we need to derive the inverse function of the CDF for the inverse transform method, we derived it using Mathematica [27]. It is written as

$$\mathrm{Cdf}^{-1}_{\mathrm{LMR}}(t) = \frac{\lambda W\left(-\frac{(1-t)^{\frac{s}{\lambda}}}{e}\right) + \lambda - s\log(1-t)}{\lambda s} \tag{9}$$

where W is the Lambert W function. We generated the random variables following the above mentioned distribution using both the random variables of the uniform distribution and the Lambert W function in SciPy.

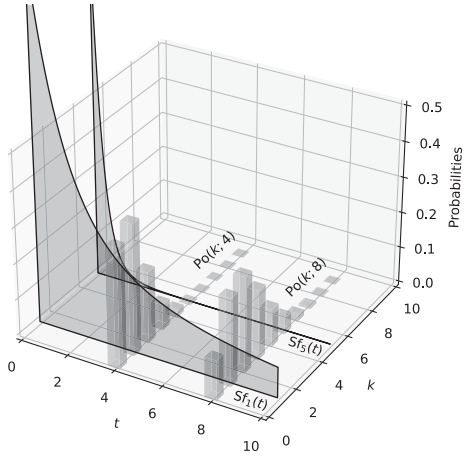

Fig. 6. Schematical view of the conditional survival functions and the Poisson process in the LMR.

It is known that the marginal distribution of Ryu's BVE was fully developed from stochastic processes considering physical behavior [28]. Then, a stochastic process exists behind the LMR presented in this article too. The stochastic process and conditional survival function in deriving the distribution match the growth process characterized by t in training. It is discussed in the next section.

IV. Discussion

A. Applying the Logic of LMR to Flight Training

The meaning of t is investigated by relating t to the context of survival analysis and following the derivation of the LMR. The marginal distribution of Ryu's BVE is constructed using survival analysis. Survival analysis analyzes the expected duration until a well-defined event occurs. For the present case, the duration is t, and the well-defined event is to complete a flight. Then, the mean of t by completing a flight is interpreted as the mean time by occurring the event. The relation between the survival analysis and the flight training is schematically shown in Fig. 6.

The LMR is derived as the expectation of conditional survival functions according to the proof in the first proposition of Appendix A [26]. The conditional survival function is written as

$$\mathrm{Sf}_k(t) = \left(\frac{1 - e^{-st}}{st}\right)^k. \tag{10}$$

This model shows that t probabilistically decreases as k increases. Then, we consider k as the training step. One step is determined for any given flight training. The stochastic process of k is modeled as a Poisson process for the number of flags. It models that the training step will be higher if the more flags are tallied. It is a natural consideration when the applicant grows by experiencing the event the flag tallied. Considering the variation in t regardless of the step, we can derive the survival function of the LMR as the expectation of the conditional survival function

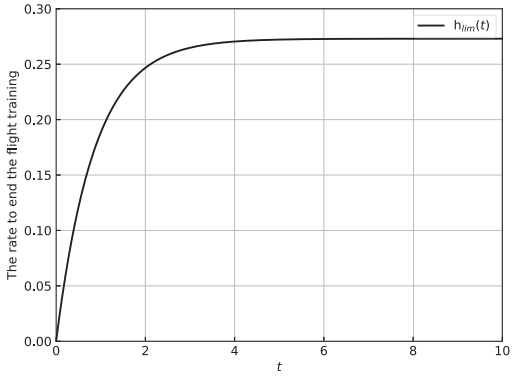

Fig. 7. Hazard function of the LMR.

for k

$$\mathrm{Sf}_{\mathrm{LMR}}(t) = \sum_{k=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \left(\frac{1 - e^{-st}}{st} \right)^k. \tag{11}$$

We consider at least two reasons why LMR and compound gamma distribution fit well. One is that both distributions are derived as compounding of two probability distributions, and the other is that one of the compounded distributions is the gamma distribution. The PDF of the compound gamma distribution is derived as

$$\mathrm{pdf}_{\mathrm{cg}}(t; \alpha, \beta, q) = \int_0^{\infty} \mathrm{pdf}_{\mathrm{gam}}(t; \alpha, r) \mathrm{pdf}_{\mathrm{gam}}(r; \beta, q) \mathrm{d}r. \tag{12}$$

The survival function of the LMR is written as (11). The Poisson distribution of k can be interpreted as a gamma distribution of t. Then, both distributions are derived as compounding of the gamma distribution and the other distribution. As the fact that they are expected values of the probability distributions means that there are two probability distributions, both the variation in individual competency and the variation in the number of flags tallied during training can be modeled. Also, the GoF of the gamma distribution was not bad. Then, we consider that the gamma distribution included in the integrand gave the better GoF.

B. Range of the Number of Flags to Keep

The hazard function of the LMR is written as

$$\mathrm{h}_{\mathrm{LMR}}(t) = \lambda \left(1 - e^{-st}\right). \tag{13}$$

It starts from 0 at $t = 0$ and rapidly converges to λ as $t \to \infty$. It is shown in Fig. 7. For example, the instantaneous rate at $t = 5$ is about 0.997λ. It shows that the rate to end the flight varies with a couple of flags, but little or no change is seen with more flags. Then, the training data indicated that keeping t within a few is important.

C. Derivation of a Growth Model Using LMR

The evolution of t against N is an important indicator of an applicant's proficiency since keeping t low is important. The concept of SRGM is considered as an analogy in the present training. Then, we will discuss what can be learned from the LMR and what is desired in analyzing the personal growth process.

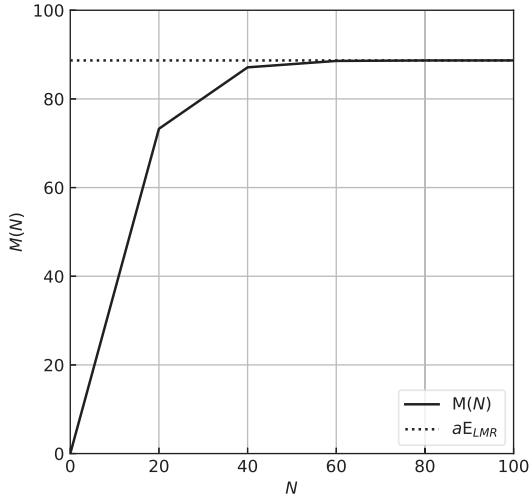

Fig. 8. $\mathrm{M}(N)$ for $a = 20$.

The expectation can be calculated by integrating the survival function. The expectation of the LMR is given as

$$\mathrm{E}_{\mathrm{LMR}} = \int_0^{\infty} \sum_{k=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \left(\frac{1 - e^{-st}}{st} \right)^k \mathrm{d}t. \tag{14}$$

When the internal function can be written in terms of power series, we can change the order of the integral and the infinite series:

$$\mathrm{E}(k) = \int_0^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \left(\frac{1 - e^{-st}}{st} \right)^k \mathrm{d}t \tag{15a}$$

$$\mathrm{E}_{\mathrm{LMR}} = \sum_{k=0}^{\infty} \mathrm{E}(k). \tag{15b}$$

They have closed forms as

$$\mathrm{E}(k) = \left(\frac{\lambda}{s} \right)^k \frac{\Gamma\left(\frac{\lambda}{s}\right)}{s\Gamma(k + \frac{\lambda}{s} + 1)} \tag{16a}$$

$$\mathrm{E}_{\mathrm{LMR}} = \frac{1}{s} \left(\frac{\lambda}{s} \right)^{1 - \frac{\lambda}{s}} e^{\frac{\lambda}{s}} \frac{\Gamma\left(\frac{\lambda}{s}\right) \left(\Gamma\left(\frac{\lambda}{s}\right) - \Gamma_i\left(\frac{\lambda}{s}, \frac{\lambda}{s}\right) \right)}{\Gamma\left(\frac{\lambda}{s} + 1\right)}. \tag{16b}$$

$\mathrm{E}(k)$ is the ratio of $\mathrm{E}_{\mathrm{LMR}}$ to the flight training at the kth step. The abovementioned model does not include how many flights it takes to reach the $(k+1)$th step of flight. Introducing a parameter, a, which is the required time of flights to reach the next step, the difference equation of the total number of flags is written as

$$\mathrm{M}(N+1) - \mathrm{M}(N) = \mathrm{E}\big(K(N)\big) \tag{17a}$$

$$K(N) = \left\lfloor \frac{N}{a} \right\rfloor \tag{17b}$$

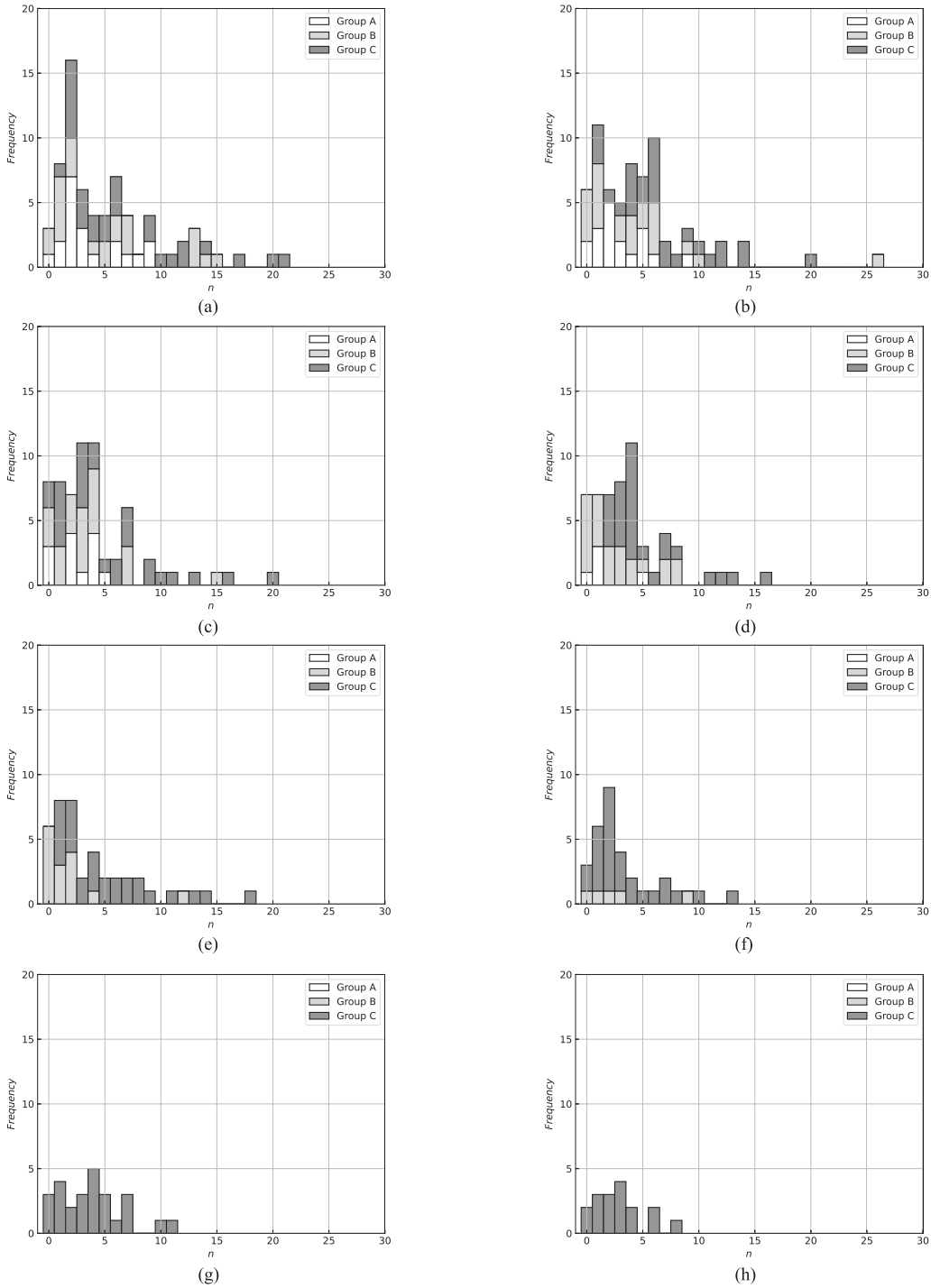

Fig. 9. Histograms of n at (a) 1st, (b) 10th, (c) 20th, (d) 40th, (e) 60th, (f) 80th, (g) 100th, and (h) 120th flights.

where $K(N)$ is a function that returns the step for the number of flights. $\mathrm{M}(N)$ is written as

$$\mathrm{M}(N) = a \sum_{i=0}^{K(N)} \mathrm{E}(i) + \Big(N - \big(K(N)+1\big)a\Big)\mathrm{E}\big(K(N)\big). \tag{18}$$

We can consider $\mathrm{M}(N)$ as a growth model that converges to $a\mathrm{E}_{\mathrm{LMR}}$. An example for $a = 20$ is shown in Fig. 8. It shows a mean growth model of the cumulative number of flags to the times of flight in a concave shape. Also, the tendency for most flags to be tallied in early training and almost none as training progresses is a reasonable model for an applicant's progress toward proficiency.

Although (18) is a line graph and a poor model for fitting, we consider it is a conceptual growth process for an applicant in standard progress. The probabilistic distribution obtained from Sf_k depends on the training step, k, which increments as

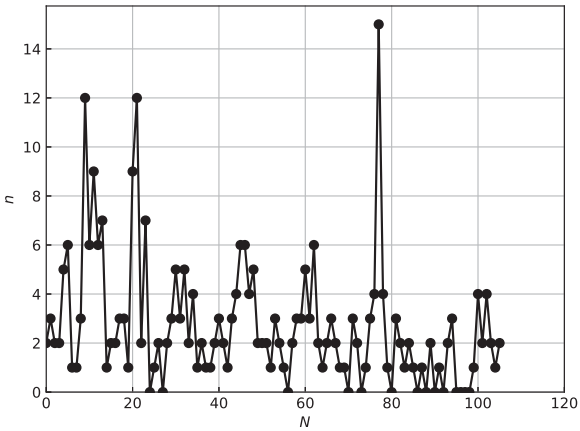

Fig. 10. History of n against N for an applicant.

the training progresses. It implies that the individual growth process would follow such a trajectory if the training continues indefinitely. It also provides a support that SRGM is one of the potential approaches that can describe the individual growth model.

D. Expected Future Work

We end our discussion by considering motivations for applying SRGM to the personal growth process during flight training. We start with examining fundamental assumptions of SRGM in terms of the flight training. Next, we elucidate the property of n as a random variable by visualizing the histories of n against N. At last, we show the history of n_{cum} for each applicant and examine its shape in light of SRGM.

At first, we discuss how fundamental assumptions of the SRGM are interpreted in the pilot training. According to assumption matrix [30], the fundamental assumptions common to all SRGMs are written as follows.

1) Software testing and reliability assessment are performed in actual operating conditions.
2) Faults are removed immediately.
3) New faults are not brought in the process of debugging.
4) All faults occurred independently from each other.

Assumption 1 is valid since the flight training is in the form that an applicant in the captain's seat operates an actual flight with passengers, which is the actual operating condition itself.

Assumption 2 means that, in terms of training, all flagged performances by an instructor on the last flight will be properly handled by the next flight. Ideally, we would like this to be the case, but in practice, it may not be easy. In order to take into account this, it will be necessary to introduce a parameter, such as a fault removal efficiency [21].

Assumption 3 is interpreted as that improved actions in light of a flag do not trigger another flag. Such a trigger is sometimes observed since the management balance is important for the captain. It is possible to be too concerned about the last flag to respond to other things appropriately. This assumption is a strong limitation in software development, and models incorporating secondary faults are being studied [31]. Those models may also be required for the personal growth model in training.

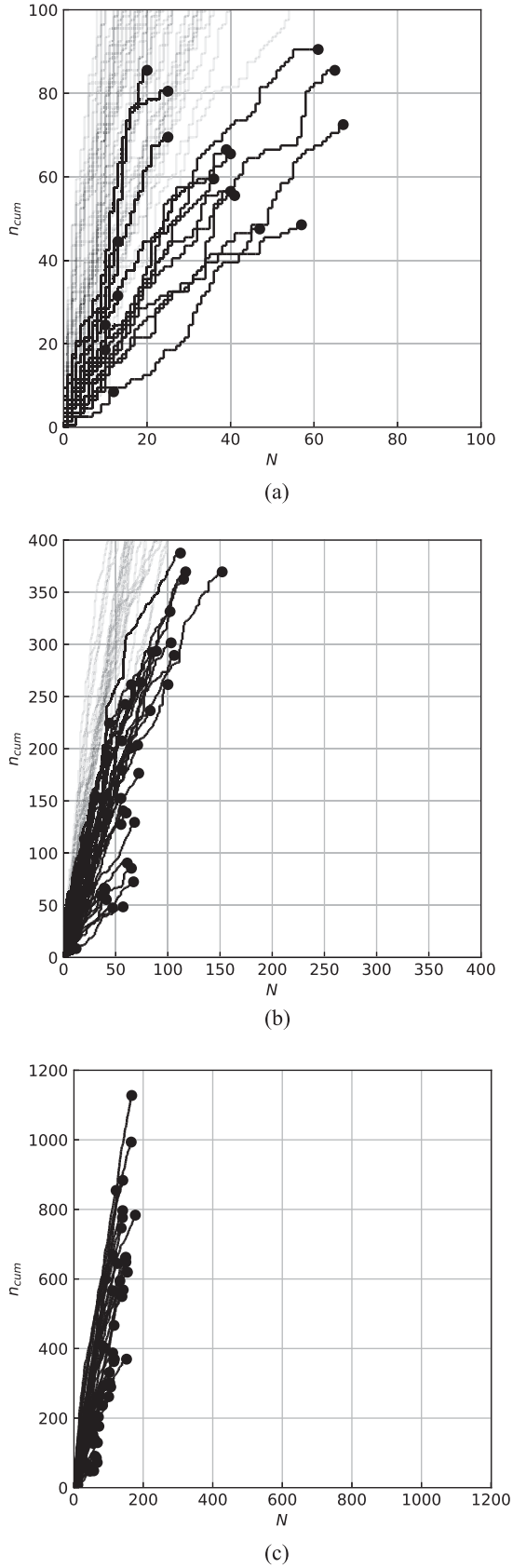

Fig. 11. Histories of n_{cum} against N as personal growth processes within (a) $N \leq 100, n_{\mathrm{cum}} \leq 100$, (b) $N \leq 400, n_{\mathrm{cum}} \leq 400$, and (c) $N \leq 1200, n_{\mathrm{cum}} \leq 1200$.

Assumption 4 is required from probability theory. It is difficult to measure the independency of events that applicants encounter during flight since they have complicated situations, such as weather, passengers, aircraft, and routes. We assume that each event independently occurs. Applicants encounter many independent events through flights with various conditions. Instructors use AMs to classify a performance or a series of performances to deal with one of the events into flags. AM is designed as a reason code that is not duplicated with each other and is extracted from ten competencies defined as dimensions of human performance by ICAO [4]. Therefore, we assume that the performance flag is independent of each other.

Next, we discuss whether the underlying assumption of personal growth processes is consistent with that of the SRGM. Applicants with $N_{\mathrm{total}} < 50$ were classified as Group A, those with $50 \leq N_{\mathrm{total}} < 100$ as Group B, and those with $N_{\mathrm{total}} \geq 100$ as Group C to investigate the trend of the personal growth process among the groups. Fig. 9 shows histograms of n tallied at the Nth flight among all applicants. We observed that n probabilistically decreased as the flight training progressed. The difference between groups was the variation in n at the first recorded flight. Group A showed that n was less than 10 from the beginning and they finished the flight training with fewer N_{total} than the other groups. Groups B and C showed that the variations in n got less at the end of training as well as Group A, but N_{total} were different.

We consider that the difference in the personal growth process among groups comes from the difference in proficiency at the entry of training. The timing to begin captain upgrade training is up to each individual. Applicants are guaranteed to be at a certain level since there is a requirement to begin captain upgrade training. Some of them are well prepared at the beginning. It is the instructor's experience that applicants' competencies are varied when they enter training but are close by the end of training. Therefore, we consider that applicants with similar N_{total} have a similar growth of proficiency.

Although the present classification is rough, the variations in n among the applicants with similar N_{total} are qualitatively visualized. We consider that the visualized variation results from the existence of a stochastic process in flight training. Fig. 10 shows the history of n against N for an applicant. n is scattered but probabilistically gets fewer as N increases. In other words, n is a random variable, and the applicant's competencies grow as fewer n is expected.

Whether an applicant adequately deals with an event encountered during flight by exerting their competencies or not is a stochastic process depending on the condition of the crews including themselves and so on, even if the applicant has learned about the event and how to deal with it in lectures or flight simulators. Since the response to the event during flight is attributed to an event in probability theory, n is a random variable generated by the stochastic process and measured as performance to be improved by an instructor. We consider the idea of growth that the applicant's performance gets stable through the accumulation of experience as n_{cum} has an analogy with the concept of the SRGM.

Finally, by analogy with the SRGM plotting the total number of errors found in the development process against time, Fig. 11 shows $n_{\mathrm{cum}}(N) \equiv \sum_{i=1}^{N} n(i)$ versus N for each applicant. The zigzag line represents the personal growth process, and the endpoint represents the point at which the training period was finished or the point at which the last flight was recorded. Plots in which the cumulative number of flags exceeds the upper limit of the axis are lightened in color. The slope of the growth process was different from each other. As the slope got lower, it seemed that N_{total} got fewer.

As the gradient of n_{cum} decreases with each training step, the convergence of the decreasing gradient is important in the growth process. However, waiting for sufficient convergence is difficult from a cost standpoint. Therefore, many processes look linear when visualized macroscopically. The findings from the simple growth model obtained by LMR capture the main points regarding modeling the training growth process. In detail, however, most look wavy concave like an s–shape [29]. Almost linear and convex processes are also observed. If the training continues, we consider these processes would be concave and s-shape processes. The end of the training is determined by the instructors. Whether the flight gets stable or not is one of the criteria. It is strongly related to the fact that n gets constant and an empirically based assumption that n will become less probable hereafter.

The above discussions suggest that the SRGM is useful in quantifying the growth process of applicants in captain upgrade training. Whether or not it can be used in practice requires detailed discussions with instructors. Therefore, quantifying the individual growth process through the application of SRGM will be the most recent future work. However, it also suggests that the instructor's assessment method may differ from that of the SRGM since the SRGM assesses the reliability of software by the total errors removed so far instead of the errors removed per day most recently. It is not yet determined what model will best represent the growth process, but the model is expected to be constructed to satisfy the requirements to improve the training.

V. Conclusion

This article clarifies the probabilistic nature of the number of flags for an applicant in standard progress by applying basic methods of reliability theory to the number of flags observed as the difference relative to the ideal competencies in CBCT developed by JAL.

First, we fit a histogram of the raw data, excluding outliers. We have identified the limit of the marginal distribution of Ryu's BVE (LMR) that best fits the probability distribution of the expected value per flight, which was the differential component of the growth process. The random variable that made the number of flags smaller as it got larger was modeled as the training step by the derivation process of the LMR.

Second, the hazard function of the LMR showed that it increased for a couple of flags per flight but remained almost unchanged for higher numbers of flags. The training data indicated that keeping the number of flags per flight to a few was important.

Next, the occupation ratio of the expected number of flags per flight was calculated for each training step. It gave the gradient of the cumulative number of flags at each training step from the viewpoint of growth models. Introducing the times of flights to progress to the next training step gave a concave growth model, which was natural for an applicant in standard progress that the many flags were tallied at the beginning of training while fewer flags were tallied as the training progressed.

At last, motivations for applying SRGM were discussed from the viewpoints of underlying assumptions and shapes of trajectories of personal growth processes. We confirmed that the fundamental assumptions of SRGM were applicable to the flight training. When some assumptions were not valid, parameters to relax the limitations would be required. Next, the transition of the number of flags per flight with the number of flights was investigated by classifying the applicants into three groups by the total number of flights. The variation in the number of flags at the entry of training was different among the groups. The variations probabilistically decreased as the training progressed and they got similar to each other at the end of training. We attributed the cause of the difference in the total number of flights to the applicant's proficiency at the entry of training. Regarding the variation in the number of flags, we attributed the cause of the variation to the existence of a stochastic process. The stochastic process was whether an applicant adequately deals with an event encountered during flight or not. Then, the number of performance flags was a random variable generated by the stochastic process. The underlying assumption of growth in flight training was that the applicant's performance gets stable through the accumulation of experience as the number of flags. Consequently, we considered that the assumption of the personal growth process matched that of the SRGM. Also, visualization of personal processes revealed that most processes were macroscopically linear but were concave, s-shaped, and in their earlier forms in detail. It was suggested that the personal growth processes ended up on the way of convergence due to practical aspects such as training cost. Then, applying the SRGM to the personal processes is expected to quantify the cumulative number of flags in total and how much percentage is tallied.

The fundamental analysis of reliability theory provided useful insights into the gross nature of flight training. However, the personal growth process is not investigated in detail. Applying various models including the SRGM to the individual growth processes is expected in the future.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for improving our paper.

REFERENCES

[1] European Union Aviation Safety Agency "Commission Regulation (EC) No1056/2008," *Official J. European Union* OJ L 283, pp. 5–29 2008.

[2] Federal Aviation Administration, "Advanced Qualification Program," AC 120-54A CHG 1, 2017. [Online]. Available: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-54A_CHG_1.pdf

[3] International Civil Aviation Organization, *Manual of Evidence-based Training*, 1 ed., Doc 9995-AN/497, 2013.

[4] International Civil Aviation Organization, *Procedures for Air Navigation Services*, 2 ed., Doc 9868, 2016.

[5] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 70–73, Jul./Aug. 2006.

[6] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 2: A macrocognitive model," *IEEE Intell. Syst.*, vol. 21, no. 5, pp. 88–92, Sep./Oct. 2006.

[7] G. Klein, J. Phillips, E. L. Rall, and D. A. Peluso, "A data-frame theory of sensemaking," in *Proc. 6th Int. Conf. Naturalistic Decis. Mak.*, 2007, pp. 113–155, doi: 10.4324/9780203810088.

[8] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, "Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise," *Hum. Factors: J. HFES*, vol. 59, pp. 1161–1172, 2017, doi: 10.1177/0018720817723428.

[9] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, "The influence of surprise on upset recovery performance in airline pilots," *Int. J. Aerosp. Psychol.*, vol. 27, no. 1/2, pp. 2–14, 2017, doi: 10.1080/10508414.2017.1365610.

[10] J. A. Caldwell Jr, "Fatigue in the aviation environment: An overview of the causes and effects as well as recommended countermeasures," *Aviation, Space, Enviromental Med.*, vol. 68, pp. 932–938, 1997.

[11] N. B. Sarter, R. J. Mumaw, and C. D. Wickens, "Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data," *Hum. Factors: J. HFES*, vol. 49, no. 3, pp. 347–357, 2007, doi: 10.1518/001872007X196685.

[12] S. M. Casner, R. W. Geven, and K. T. Williams, "The effectiveness of airline pilot training for abnormal events," *Hum. Factors: J. HFES*, vol. 55, pp. 477–485, 2013, doi: 10.1177/0018720812466893.

[13] A. Landman, P. van Oorschot, M. M. van Paassen, E. L. Groen, A. W. Bronkhorst, and M. Mulder, "Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios," *Hum. Factors: J. HFES*, vol. 60, pp. 793–805, 2018, doi: 10.1177/0018720818779928.

[14] K. W. Lee, J. J. Higgins, and F. A. Tillman, "Stochastic modelling of human-performance reliability," *IEEE Trans. Rel.*, vol. 37, no. 5, pp. 501–504, Dec. 1988, doi: 10.1109/24.9871.

[15] K. W. Lee, J. J. Higgins, and F. A. Tillman, "Stochastic models for mission effeciveness," *IEEE Trans. Rel.*, vol. 39, no. 3, pp. 321–324, Aug. 1990, doi: 10.1109/24.103011.

[16] K. W. Lee, "Stochastic models for random-request availability," *IEEE Trans. Rel.*, vol. 49, no. 1, pp. 80–84, Mar. 2000, doi: 10.1109/24.855539.

[17] L. Ciani, G. Guidi, G. Patrizi, and D. Galar, "Improving human reliability analysis for railway systems using fuzzy logic," *IEEE Access*, vol. 9, pp. 128648–128662, 2021, doi: 10.1109/ACCESS.2021.3112527.

[18] D. Wang, Y. Wei, J. Zhan, L. Xu, and Q. Lin, "Human reliability assessment of home-based rehabilitation," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1310–1320, Dec. 2021, doi: 10.1109/TR.2020.3001923.

[19] W. Q. Meeker, L. A. Escobar, and F. G. Pascual, *Statisitical Methods for Reliability Data*, 2nd ed., Hoboken, NJ, USA: Wiley, 2021.

[20] A. L. Goel and K. Okumoto, "Time-dependent error-detection rate model for software reliability and other performance measures," *IEEE Trans. Rel.*, vol. R-28, no. 3, pp. 206–211, Aug. 1979, doi: 10.1109/TR.1984.5221826.

[21] X. Zhang, X. Teng, and H. Pham, "Considering fault removal efficiency in software reliability assessment," *IEEE Trans. Syst., Man, Cybern. - Part A, Syst. Hum.*, vol. 33, no. 1, pp. 114–120, Jan. 2003, doi: 10.1109/TSMCA.2003.812597.

[22] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2001. [Online]. Available: http://www.scipy.org/

[23] W. Stute, W. G. Manteiga, and M. P. Quindmil, "Bootstrap based goodness-of-fit-tests," *Metrika*, vol. 40, pp. 243–256, 1993, doi: 10.1007/BF02613687.

[24] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton Univ. Press, 1946.

[25] M. J. De Smith, *Statistical Handbook: A Comprehensive Guide to Statistical Concepts Methods and Tools*, London, U.K.: Drumlin Secur. Ltd., 2018.

[26] K. Ryu, "An extension of marshall and Olkin's bivariate exponential distribution," *J. Amer. Statist. Assoc.*, vol. 88, no. 424, pp. 1458–1465, 1993, doi: 10.1080/01621459.1993.10476434.

[27] *Mathematica, Version 12.3.1*, Wolfram Research Inc. [Online]. Available: https://www.wolfram.com/mathematica

[28] M. G. Edra, *Properties of Order Statistics From Bivariate Exponential Distributions*. Ann Arbor, MI, USA: Ohio State Univ., 1994.
[29] S. Yamada, M. Ohba, and S. Osaki, "s-shaped software reliability growth models and their applications," *IEEE Trans. Rel.*, vol. R-33, no. 5, pp. 289–292, Oct. 1984, doi: 10.1109/TR.1984.5221826.
[30] V. S. Kharchenko, O. M. Tarasyuk, V. V. Sklyar, and V. Y. Dubnitsky, "The method of software reliability growth models choice using assumptions matrix," in *Proc. 26th Annu. Int. Comput. Softw. Appl.*, 2002, pp. 541–546, doi: 10.1109/CMPSAC.2002.1045062.
[31] P. Zeephongsekul, G. Xia, and S. Kumar, "Software-reliability growth model: Primary-failures generate secondary-faults under imperfect debugging," *IEEE Trans. Rel.*, vol. 43, no. 3, pp. 408–413, Sep. 1994, doi: 10.1109/24.326435.

Kento Yamada received the Ph.D. degree in engineering from the University of Tokyo, Tokyo, Japan, in 2019.

He is currently a Researcher with the Aviation Technology Directorate, JAXA, Tokyo.

Harumi Ikeshita received the M.E. degree in applied system science from Kyoto University, Kyoto, Japan, in 1991.

He is currently an Administrator of Pilot training planning for JAL, Tokyo, Japan.

Yuta Kyoya received the B.E. degree in mechanical engineering from the Tokyo University of Science, Tokyo, Japan, in 1992.

He is currently a Boeing 767 Captain and a Flight Training Instructor for JAL, Tokyo.

Makoto Ueno received the Doctor of Engineering degree in aerospace engineering from Nagoya University, Nagoya, Japan, in 2017.

He is currently a Senior Researcher with the Aviation Technology Directorate, JAXA, Tokyo, Japan.