An Exploration of a Testing Procedure for the Aviation Industry

by

Liwei Lai

Bachelor of Science (Hons.), University of Manitoba, 2020

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2023

Declaration of Committee

Name:	**Liwei Lai**
Degree:	**Master of Science**
Title:	**An Exploration of a Testing Procedure for the Aviation Industry**
Committee:	**Chair: Liangliang Wang** Associate Professor, Statistics and Actuarial Science
	Tim Swartz Co-Supervisor Professor, Statistics and Actuarial Science
	Gary Parker Co-Supervisor Associate Professor, Statistics and Actuarial Science
	Joan Hu Committee Member Professor, Statistics and Actuarial Science
	Sonja Isberg Examiner Lecturer, Statistics and Actuarial Science

Abstract

In the aviation industry, pilot training is paramount, necessitating robust and precise assessment methodologies. Despite the shift towards Competency-based Training and Assessment (CBTA) recommended by the International Civil Aviation Organization (ICAO), there is a notable absence of comprehensive statistical models to substantiate the evaluation process. This project explores the application of an enhanced Many-Facet Rasch Model (MFRM) employing Bayesian estimation techniques and presents a novel approach for quantifying pilot competency scores, ensuring a more granular and accurate assessment of pilot capabilities. By analyzing simulated data, the research assesses the viability of this statistical approach in operational settings. Potential applications and limitations of this methodology within the aviation industry are discussed.

Keywords: Many-Facet Rasch Model; Bayesian Estimation; Competency-based Training and Assessment; Stan

Acknowledgements

I would like to extend my deepest gratitude to several key individuals and organizations who have played an integral role in my academic journey and the completion of this project.

First and foremost, I am profoundly grateful to my parents, whose unwavering support and encouragement have been my guiding light throughout this endeavor. Their belief in me and my capabilities has been a source of immense strength and motivation.

I am immensely thankful to my supervisors, Dr. Tim Swartz and Dr. Gary Parker, whose guidance, expertise, and patience have been fundamental in my research journey. Their constructive feedback, encouragement, and dedication have been pivotal in the successful completion of this project.

Special thanks go to my uncle, Yuping Guan, who provided the initial idea for this project. His extensive experience in the aviation industry offered unique insights and perspectives that greatly enriched my research.

My appreciation extends to Astrom Training Solutions, a company that provides training solutions for airline companies worldwide. Their cooperation and willingness to share detailed information about the aviation industry significantly enhanced my understanding and brought a global perspective to my theoretical research.

I also want to express my heartfelt appreciation to my friends. Their constant encouragement, understanding, and companionship provided much-needed relief and support during the challenging moments of this journey.

To all these individuals and organizations, I express my heartfelt gratitude. Your contributions have been instrumental in my academic and personal growth, and I am deeply thankful for your involvement in my journey.

Table of Contents

List of Tables

Chapter 1.

Introduction

Civil aviation is an integral component of transportation within our modern society, serving as one of the primary modes for long-distance travel for most individuals. As highly sophisticated products of modern technology, airlines necessitate stringent qualifications from their pilots. Consequently, pilot training emerges as a crucial element in the progression of the civil aviation industry.

Pilots are obligated to complete recurrent training regularly on simulators to maintain their certification. In traditional qualification-based trainings, the examination is divided into different sections corresponding to different flight phases (take-off, climb, cruise, etc.). For each phase, pilots are assigned various tasks and receive a score from the instructor for each task. The scoring system typically employs a five-point scale, with 1 as the lowest and 5 as the highest score. Some countries, however, use a four-point scale in their aviation pilot training, but in both systems, a score of 3 is considered passing. The aim of the qualification-based training is to ascertain that pilots maintain the essential knowledge, skills and experience required to uphold the standards of their license.

With advancements in aircraft technology and growing demands for more effective pilot evaluation methods, the International Commercial Aviation Organization (ICAO) has, since 2020, introduced a competency-based training and assessment approach. A competency is a combination of knowledge, skills, and attitudes required to perform a task to the prescribed standard (ICAO 2018). The ICAO has defined 9 competencies as follows:

1. Application of Procedures (PRO)

2. Communication (COM)

3. Flight Path Management, Automation (FPA)

4. Flight Path Management, Manual (FPM)

5. Leadership & Teamwork (LTW)

6. Problem Solving & Decision Making (PSD)

7. Situation Awareness (SAW)

8. Workload Management (WLM)

9. Application of Knowledge (KNO)

The goal of competency-based training is to ensure pilots develop required competencies to carry out their assigned duties and responsibilities safely, efficiently and effectively at workplace (Defalque 2017). In the competency-based training, each task is associated with specific competencies, typically three. Rather than grading individual tasks, instructors assess each competency based on the pilot's overall performance during the training session. This method allows aviation employers to construct their own assessments under the ICAO competency framework, ensuring the content is relevant to the job's actual requirements. Furthermore, it optimizes the utilization of training tools and methodologies, leading to a more holistic and applicable assessment of pilot abilities.

At present, airlines in Europe and the Middle East rely exclusively on competency-based training, where instructors grade each competency based on a pilot’s overall performance. In contrast, airlines in the U.S. and Canada use a hybrid grading system. This system combines both task-based and competency-based assessments: during a test, instructors assign a score to each task, and after the test, they grade each competency based on the pilot’s overall performance. Meanwhile, airlines in China are transitioning towards the competency-based training system, gradually phasing out their task-based grading approach.

Despite the many advantages of competency-based training, it also presents some unresolved challenges. A prevailing issue is linking competencies and tasks during grading. Although each task associates with specific competencies, grades for each competency are determined based on overall performance, rather than being anchored to the tasks directly associated with that competency. This lack of specificity can make it

challenging to design targeted training plans for improvement. Therefore, it is crucial to retain the practice of scoring individual tasks alongside competency assessments.

Another key issue is the subjective nature of scoring. Whether in task-based or competency-based training, scores are determined by instructors, introducing variability in grading. Some instructors might be more lenient, while others could be stricter. As a result, the same pilot performance might yield different scores from different instructors, thereby challenging the reliability and consistency of the assessment process.

The primary aim of this project is to devise a method for determining suitable scores for each competency derived from the raw scores of task grading, while also mitigating the impact of instructors' subjective judgements in the assessment process. Specifically, we are interested in employing a model from Item Response Theory to accomplish our goal.

Item Response Theory (IRT), also known as Latent Trait Theory, is a testing method for designing, analyzing, and scoring tests and questionnaires. It is particularly adept at assessing abilities, attitudes, and other distinct variables. Generally, IRT attempts to mathematically model the probability of an identified response based upon the interaction of a person's latent ability and characteristics (e.g., difficulty) of an item (Ackerman et al., 2023, p.72).

The predominant use of IRT is in the field of education, where psychometricians employ it to craft and structure examinations. IRT's foundational development as a theoretical framework took place in the 1950s, when Lord published his doctoral dissertation on the theory of latent traits which led to a few other publications (Lord, 1950; Lord, 1952; Lord, 1953). Rasch (1960) released a book introducing several item response models, most notably the renowned Rasch model, which addresses dichotomous responses. Following Rasch's development of a polytomous model (1961), variations of the model were subsequently introduced. Samejima (1969) introduced the Graded Response Model. Later, Andrich (1978) formulated the Rating Scale Model, which reinterprets terms from Rasch’s model as thresholds, particularly focusing on Likert-type questions. The Partial Credit Model, introduced by Masters (1982), and the Generalized Partial Credit Model, presented by Muraki (1997), address scenarios where thresholds vary across different items.

An additional significant development within the IRT lineage is the Many-Facet Rasch Model (Linacre, 1989; Eckes, 2015). Compared to earlier polytomous models, the Many-Facet Rasch Model integrates rater severity into its analysis, alongside considerations of test-taker's ability and task difficulty. A recent advancement is the introduction of a generalized Many-Facet Rasch model that utilizes Bayesian estimation (Uto & Ueno, 2020; Uto, 2022). This model captures all the typical characteristics within the IRT framework, a feat that other IRT models couldn't achieve due to inaccurate parameter estimation resulting from model complexity. Instead of using traditional maximum likelihood estimation, Uto & Ueno (2020) adopted a robust Bayesian estimation method using the No-U-Turn Hamiltonian Monte Carlo algorithm, an MCMC algorithm.

In this project, we adopt the Bayesian Many-Facet Rasch Model and propose a reparametrized version of it for analyzing pilot competencies. We first estimate pilots' abilities, task difficulties and instructor severities based on raw task scores. Following this, we project scores for pilots across all tasks in our repository, assuming they were evaluated by an average instructor. This will allow us to deduce the score of each competency for pilots.

All computations are performed using R and Stan; the latter leverages Hamiltonian Monte Carlo techniques to produce posterior samples (Carpenter et al., 2017).

The main challenge of this project is that, while the primary interest is rooted in a genuine industry challenge, we were unable to obtain actual data from airline companies. In this project, we rely on simulation studies and in-depth discussions. In Chapter 2, we elucidate the fundamental statistical model and detail its integration within a Bayesian framework. In Chapter 3, simulation studies are undertaken, encompassing data generation and posterior estimations. The raw score data are generated based on probabilities that we believe closely mirror real-world scoring. Chapter 4 delves into the potential applications of this project in industry, explores additional topics, and discusses potential real-data scenarios. It's important to underscore that this is preliminary work and not a finalized product ready for industry implementation.

Chapter 2.

Methods

2.1. Many-Facet Rasch Model

Consider a test where questions have scores l = 1, 2, …, L. The most common form of the Many-Facet Rasch model for the i^{th} pilot, j^{th} task and k^{th} instructor is specified as the following log-odd form (Eckes 2015):

$$\ln\left(\frac{P_{ijkl}}{P_{ijk(l-1)}}\right) = \alpha_i - \gamma_j - \delta_k - \tau_l \qquad \text{for } l = 2, \ldots, \mathrm{L} \tag{1}$$

where

P_{ijkl} = probability of pilot i receiving a score of l on task j from instructor k,

α_i = ability of pilot i, i = 1,…, I

γ_j = difficulty of task j, j = 1,…, J

δ_k = severity of instructor k, k = 1,…, K

τ_l = difficulty of receiving score l relative to score $l-1$, l = 2,…, L

The parameter τ_l can be interpreted as the parameter determining the threshold for getting a score of l instead of $l-1$ given that pilot receives one of these scores. The parameter τ_l determines the location where the two scores are equally likely. The probability can then be expressed as follows:

$$P_{ijkl} = \frac{\exp[\sum_{m=1}^{l}(\alpha_i - \gamma_j - \delta_k - \tau_m)]}{\sum_{l=1}^{L}\exp[\sum_{m=1}^{l}(\alpha_i - \gamma_j - \delta_k - \tau_m)]} \quad (2)$$

or

$$P_{ijkl} = \frac{\exp[l*(\alpha_i - \gamma_j - \delta_k) - \sum_{m=1}^{l}\tau_m]}{\sum_{l=1}^{L}\exp[l*(\alpha_i - \gamma_j - \delta_k) - \sum_{m=1}^{l}\tau_m]} \quad (3)$$

However, due to the summation of the parameters in (2) or the multiplication in (3), this model lacks statistical intuitiveness. Additionally, the mixture of positive and negative parameter coefficients is not intuitive. In light of these issues, we propose an alternative model that offers clearer expression and understanding.

2.2. The Proposed Model

We consider the following model with the same number of parameters as (1):

$$\ln\left(\frac{P_{ijkl}}{P_{ijkL}}\right) = \begin{cases} \alpha_i + \gamma_j + \delta_k + \tau_l & \text{for } l = 1, 2, \dots, \mathrm{L}-1 \\ 0 & \text{for } l = \mathrm{L} \end{cases} \quad (4)$$

where τ_l is now a parameter determining the difficulty of receiving a score of l relative to score L, the base score. For model simplicity, the parameter τ_l is assumed to be the same for all tasks.

Compared to the MFRM model discussed in Section 2.1, the proposed model's log-odds form considers a score in relation to a base score, a practice widely accepted

in multinomial logistic regression. Additionally, all the parameters in this model carry positive signs, which is simpler to interpret.

From (4) we obtain:

$$P_{ijkl} = P_{ijkL} * \exp(\alpha_i + \gamma_j + \delta_k + \tau_l) \quad \text{for } l = 1, \ldots, \text{L-1} \tag{5}$$

Thus,

$$1 - P_{ijkL} = P_{ijkL} * \sum_{l=1}^{L} \exp(\alpha_i + \gamma_j + \delta_k + \tau_l) \quad \text{for } l = 1, \ldots, \text{L-1} \tag{6}$$

Substituting (6) into (5), we have:

$$P_{ijkl} = \begin{cases} \frac{\exp(\alpha_i+\gamma_j+\delta_k+\tau_l)}{1+\sum_{l=1}^{L-1}\exp(\alpha_i+\gamma_j+\delta_k+\tau_l)} & \text{for } l = 1, 2, \ldots, \text{L} - 1 \\ \frac{1}{1+\sum_{l=1}^{L-1}\exp(\alpha_i+\gamma_j+\delta_k+\tau_l)} & \text{for } l = \text{L} \end{cases} \tag{7}$$

2.3. Bayesian Formulation

For complex models, a Bayesian estimation method generally provides more robust estimations (Uto & Ueno, 2020). In the application of pilot testing, there is considerable knowledge associated with the pilots, tasks and instructors. Thus, the model can be formulated in a Bayesian framework by introducing informative priors on the model parameters. Since in our exploration we do not have access to actual pilot data, the introduction of prior information is carried out at a theoretical level.

In this project, we introduce a multivariate normal prior for the parameters. There are many advantages of choosing a multivariate normal prior. Foremost, it is trackable and it offers concave shapes that we are likely to expect when eliciting prior opinion. The normal distribution also permits an intuitive covariance structure.

Furthermore, the multivariate normal prior brings flexibility, especially when investigating possible correlated relationships. We think in the most general case, there are correlations between pilots, tasks, and instructors. For example:

1. Pilots who test well may be treated leniently by instructors.

2. An instructor could be stringent regarding certain tasks.

Correlations may also exist within parameters. For example, tasks within the same category are likely correlated. Landing during the day and landing at night should have relevance to one another.

In the context of proposed model (7), we denote the parameter vectors as:

$$\alpha = (\alpha_1, \dots, \alpha_I)$$

$$\gamma = (\gamma_1, \dots, \gamma_J)$$

$$\delta = (\delta_1, \dots, \delta_K)$$

$$\tau = (\tau_1, \dots, \tau_{L-1})$$

Then the multivariate normal prior is

$$(\alpha, \gamma, \delta, \tau)^T \sim N_{I+J+K+(L-1)}(\mu, \Sigma)$$

where

$$\mu = (\mu_\alpha, \mu_\gamma, \mu_\delta, \mu_\tau)^T$$

$$\mu_\alpha = (\mu_{\alpha_1}, \dots, \mu_{\alpha_I})$$

$$\mu_\gamma = (\mu_{\gamma_1}, \dots, \mu_{\gamma_J})$$

$$\mu_\delta = (\mu_{\delta_1}, \dots, \mu_{\delta_K})$$

$$\mu_\tau = (\mu_{\tau_1}, \dots, \mu_{\tau_{L-1}})$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{\alpha\alpha} & \Sigma_{\alpha\gamma} & \Sigma_{\alpha\delta} & \Sigma_{\alpha\tau} \\ \Sigma_{\alpha\gamma}{}^T & \Sigma_{\gamma\gamma} & \Sigma_{\gamma\delta} & \Sigma_{\gamma\tau} \\ \Sigma_{\alpha\delta}{}^T & \Sigma_{\gamma\delta}{}^T & \Sigma_{\delta\delta} & \Sigma_{\delta\tau} \\ \Sigma_{\alpha\tau}{}^T & \Sigma_{\gamma\tau}{}^T & \Sigma_{\delta\tau}{}^T & \Sigma_{\tau\tau} \end{pmatrix}$$

Each component of the variance-covariance matrix is a sub-matrix. For example, $\Sigma_{\alpha\alpha}$ is the variance matrix of parameter α, and $\Sigma_{\alpha\gamma}$ is the covariance matrix of α and γ.

Intuitively, on the diagonal, $\Sigma_{\alpha\alpha}$ and $\Sigma_{\delta\delta}$ are diagonal matrices taking the form $\sigma_\alpha^2 I$ and $\sigma_\gamma^2 I$, respectively because we think pilots are independent individuals as are instructors. Among the remaining matrices, there are unlikely correlations between relative difficulty and other parameters, so $\Sigma_{\tau\tau}$, $\Sigma_{\alpha\tau}$, $\Sigma_{\gamma\tau}$ and $\Sigma_{\delta\tau}$ should be 0. Hence, the focus of the variance-covariance matrix falls on $\Sigma_{\gamma\gamma}$, $\Sigma_{\alpha\gamma}$, $\Sigma_{\alpha\delta}$ and $\Sigma_{\gamma\delta}$, i.e., within tasks, between pilots and tasks, between pilots and instructors, and between tasks and instructors.

2.4. Devising the Competency Scores

After obtaining the estimated parameters for pilots, tasks, instructors and relative difficulties using a Bayesian approach, we proceed to calculate a score for each of the nine competencies for every pilot. This involves two steps.

First, we project the scores for each pilot on every task, assuming they are graded by an instructor of average severity. To achieve this, we insert the estimated pilot

ability and task difficulty into model (7) for each pilot and each task. We then use a value of 0 for instructor severity. Thus, for pilot i and task j, the estimated probability of getting score l is

$$\hat{P}_{ijl} = \begin{cases} \frac{\exp(\hat{\alpha}_i+\hat{\gamma}_j+\hat{\tau}_l)}{1+\sum_{l=1}^{L-1}\exp(\hat{\alpha}_i+\hat{\gamma}_j+\hat{\tau}_l)} & \text{for } l = 1, 2, \dots, \mathrm{L}-1 \\ \frac{1}{1+\sum_{l=1}^{L-1}\exp(\hat{\alpha}_i+\hat{\gamma}_j+\hat{\tau}_l)} & \text{for } l = \mathrm{L} \end{cases} \quad (8)$$

Thus, the projected score can be calculated by

$$S_{ij} = \sum_{l=1}^{L} \hat{P}_{ijl} * l \quad (9)$$

For pilot i, we define the projected scores as:

$$S_i = (S_{i1,} S_{i2,} \dots, S_{iJ})$$

Second, as previously mentioned, each task is associated with specific competencies. We think that different tasks have different weights for competencies. For example, a task associated with competencies of communication, situation awareness, and leadership might have weights of 40%, 40% and 20%, respectively. In contrast, another task associated with those same competencies could have weights of 30%, 30% and 40%, respectively. Thus, for task j, we define the competency weights as:

$$w_j = (w_{j1}, w_{j2}, w_{j3}, w_{j4}, w_{j5}, w_{j6}, w_{j7}, w_{j8}, w_{j9})$$

for the nine competencies. The weights are non-negative, $w_{jc} \geq 0$. They sum to 1, $\sum_{c=1}^{9} w_{jc} = 1$, and only the weights of the associated competencies are non-zero. The weights should be determined by experts in the aviation industry.

With the projected task scores and competency weights of each task, we calculate the score for a competency by taking the weighted average of the scores of tasks associated with this competency. Therefore, for pilot i, the competency score on competency c is:

$$\theta_{ic} = \frac{\sum_{j=1}^{J}(w_{jc}*S_{ij})}{\sum_{j=1}^{J} w_{jc}} \text{ for i = 1, 2, …, } I \text{ and c = 1, 2, …, 9} \quad (10)$$

Chapter 3.

Simulation Study

3.1. Model Validation

In this subsection, we carry out a simulation to confirm that the Bayesian implementation can accurately estimate the underlying parameters.

For this purpose, we create a simple dataset consisting of scores for 6 pilots, 4 tasks and 3 instructors. All the tasks are associated with a particular competency. The task scores are on a 5-point scale, with score 5 being the highest and 1 the lowest. To construct the dataset, we first assign parameter values for model (7) as detailed in Table 1 to pilots, tasks and instructors.

	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
Pilot ability α_i	-0.8	-0.3	0.0	0.2	0.4	1.0
Task difficulty γ_j	0.8	0.0	-0.3	-0.5		
Instructor severity δ_k	1.0	0.0	-1.0			

Table 1 The assigned parameter values for the generated dataset.

Pilots with higher ability values are deemed as proficient; tasks with greater positive parameter values are considered easy; and instructors with higher parameter values are seen as lenient.

For the relative difficulty τ_l, score 1 is set as the base score, so τ_1 is 0. The assigned values are as follows:

Score	1	2	3	4	5
Relative difficulty τ_l	0	-1	-0.5	1	1.5

Table 2 Relative difficulty for generated dataset.

Given the assigned parameters, we randomly generate score data based on the model described in (7). The scores are generated for every pilot performing each task 100 times under the supervision of each instructor, resulting in a total 7,200 scores (6*4*3*100) in the dataset. Each score is generated independently. The 100 repeated tasks done by pilots are not realistic, but they are carried out so that we have strong information concerning the parameters. Table 3 presents a sample showcasing the structure of the generated data.

Pilot id	Task id	Instructor id	Score
1	1	1	3
1	1	1	4
1	2	1	4
1	2	1	5
1	2	2	1
1	2	2	2

Table 3 Sample of generated data.

With this dataset, we estimate parameters using Stan to evaluate parameter recovery. For the prior distributions, we refer to Section 2.3 and operate under the assumption that all parameters are independent, and each follows a normal distribution with mean 0 and variance 1.

Tables 4 - 7 display the result of the recovery test. Values within brackets represent the standard error of the estimated parameter. The results indicates that, with large dataset, the Bayesian implementation can accurately recover the selected parameters with small standard errors. This provides us with confidence that the Stan programming is working correctly.

	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
Pilot ability $\hat{\alpha}_i$	-0.79	-0.27	0.17	0.17	0.19	0.86
S.E.	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table 4 Estimated pilot abilities, with standard errors in parentheses.

	1st	2nd	3rd	4th
Task difficulty $\hat{\gamma}_j$	0.78	0.15	-0.22	-0.33
S.E.	(0.01)	(0.01)	(0.01)	(0.01)

Table 5 Estimated task difficulties, with standard errors in parentheses.

	1st	2nd	3rd
Instructor severity $\hat{\delta}_k$	1.22	0.12	-1.04
S.E.	(0.01)	(0.01)	(0.01)

Table 6 Estimated instructor severities, with standard errors in parentheses.

Score	1	2	3	4	5
Relative difficulty $\hat{\tau}_l$	0	-1.23	-0.63	0.89	1.33
S.E.		(0.01)	(0.01)	(0.01)	(0.01)

Table 7 Estimated relative difficulties, with standard errors in parentheses.

3.2. Model Capability

In this section we consider what we believe to be some realistic probabilities P_{ijkl}. We then investigate the resulting estimated parameters and observe whether the induced probabilities are similar to the manually selected probabilities.

While maintaining the same number of pilots, tasks, and instructors as the dataset presented in Section 3.1, we select the probability distribution for the scores of each of the 72 combinations. Table 8 displays the labels assigned to each parameter. For pilots, a higher label number signifies better proficiency. In the case of tasks, a higher label number denotes easier tasks. Lastly, a higher label number for instructors suggests greater leniency in their assessment.

Based on our knowledge, in a 5-point scale rating, a score of 3 and above is deemed as passing. Receiving a score of 1 is highly improbable, while a score of 2 is uncommon but still possible. Most pilots typically achieve scores of 3 or 4 based on their performance. A score of 5, which denotes perfection, is also quite rare. Table 9 shows part of the selected probability distributions. As observed, when both the task difficulty and instructor severity remain constant, an increase in pilot ability results in a decreased

likelihood of attaining lower scores and an increased likelihood of receiving higher scores.

Parameter	*Label*					
Pilot	1st	2nd	3rd	4th	5th	6th
Task	1st	2nd	3rd	4th		
Instructor	1st	2nd	3rd			

Table 8 The labels of pilots, tasks and instructors.

Scores are generated in a manner consistent with the approach outlined in Section 3.1. For each of the 72 pilot-task-instructor combinations, scores are independently generated based on the probability distribution. Again, while this method may not reflect a real-world scenario, it embeds strong information within the data, enabling a deeper investigation of the parameters. As in Section 3.1, we estimate the parameters using Stan, assuming all parameters are a priori independent, and each follows a normal distribution with mean 0 and variance 1. Tables 10 - 13 display the estimated parameters along with their standard errors.

Parameter labels			*Manually selected probability*				
Pilot	Task	Instructor	1	2	3	4	5
1st	2nd	2nd	0.02	0.18	0.67	0.13	0.00
2nd	2nd	2nd	0.02	0.12	0.61	0.25	0.00
3rd	2nd	2nd	0.01	0.05	0.60	0.30	0.04
4th	2nd	2nd	0.01	0.04	0.55	0.35	0.05
5th	2nd	2nd	0.00	0.02	0.5	0.41	0.07
6th	2nd	2nd	0.00	0.01	0.45	0.45	0.09
1st	3rd	2nd	0.02	0.12	0.61	0.25	0.00
2nd	3rd	2nd	0.01	0.05	0.60	0.30	0.04
3rd	3rd	2nd	0.01	0.04	0.55	0.35	0.05
4th	3rd	2nd	0.00	0.02	0.50	0.41	0.07
5th	3rd	2nd	0.00	0.01	0.45	0.45	0.09
6th	3rd	2nd	0.00	0.01	0.37	0.51	0.11

Table 9 Sample of manually selected probability distributions for the five score outcomes.

	1st	*2nd*	*3rd*	*4th*	*5th*	*6th*
Pilot ability $\hat{\alpha}_i$	-0.69	-0.27	0.16	0.84	1.14	2.13
S.E.	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table 10 Estimated pilot ability, with standard errors in parentheses.

	1st	*2nd*	*3rd*	*4th*
Task difficulty $\hat{\gamma}_j$	0.46	0.38	0.74	1.80
S.E.	(0.01)	(0.01)	(0.01)	(0.01)

Table 11 Estimated task difficulty, with standard errors in parentheses.

	1st	*2nd*	*3rd*
Instructor severity $\hat{\delta}_k$	0.49	1.15	1.73
S.E.	(0.01)	(0.01)	(0.01)

Table 12 Estimated instructor severity, with standard errors in parentheses.

Score	*1*	*2*	*3*	*4*	*5*
Relative difficulty $\hat{\tau}_l$	0	-0.07	2.06	1.66	-0.22
S.E.		(0.01)	(0.01)	(0.01)	(0.01)

Table 13 Estimated relative difficulty, with standard errors in parentheses.

Lacking the "true parameters" for direct comparison, we input the estimated parameters from Tables 10 -13 in model (7) to determine the estimated probability distributions. In Table 14, an extension of Table 9, we compare these estimated distributions with the manually selected probability distributions based on the parameter labels. It's evident that differences exist, indicating that the estimated parameters don't accurately induce the manually selected probabilities. A striking observation is the static nature of score probabilities even as pilot ability increases. This indicates that the model is incapable of identifying all potential probability distributions.

Parameter labels			*Manually selected probability*					*Estimated probability*					*Manually selected - Estimated*				
Pilot	Task	Instructor	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1st	2nd	2nd	0.02	0.18	0.67	0.13	0.00	0.02	0.06	0.53	0.34	0.05	0.00	0.12	0.14	-0.21	-0.05
2nd	2nd	2nd	0.02	0.12	0.61	0.25	0.00	0.02	0.06	0.53	0.34	0.05	0.00	0.06	0.08	-0.09	-0.05
3rd	2nd	2nd	0.01	0.05	0.60	0.30	0.04	0.01	0.06	0.53	0.34	0.05	0.00	-0.01	0.07	-0.04	-0.01
4th	2nd	2nd	0.01	0.04	0.55	0.35	0.05	0.01	0.06	0.53	0.34	0.05	0.00	-0.02	0.02	0.01	0.00
5th	2nd	2nd	0.00	0.02	0.50	0.41	0.07	0.01	0.06	0.53	0.34	0.05	-0.01	-0.04	-0.03	0.07	0.02
6th	2nd	2nd	0.00	0.01	0.45	0.45	0.09	0.01	0.06	0.53	0.34	0.05	-0.01	-0.05	-0.08	0.11	0.04
1st	3rd	2nd	0.02	0.12	0.61	0.25	0.00	0.02	0.06	0.53	0.34	0.05	0.00	0.06	0.08	-0.09	-0.05
2nd	3rd	2nd	0.01	0.05	0.60	0.30	0.04	0.02	0.06	0.52	0.34	0.05	-0.01	-0.01	0.08	-0.04	-0.01
3rd	3rd	2nd	0.01	0.04	0.55	0.35	0.05	0.01	0.06	0.53	0.34	0.05	0.00	-0.02	0.02	0.01	0.00
4th	3rd	2nd	0.00	0.02	0.50	0.41	0.07	0.01	0.06	0.53	0.34	0.05	-0.01	-0.04	-0.03	0.07	0.02
5th	3rd	2nd	0.00	0.01	0.45	0.45	0.09	0.01	0.06	0.53	0.34	0.05	-0.01	-0.05	-0.08	0.11	0.04
6th	3rd	2nd	0.00	0.01	0.37	0.51	0.11	0.01	0.06	0.53	0.34	0.05	-0.01	-0.05	-0.16	0.17	0.06

Table 14 Sample of manually selected probabilities subtracting estimated probabilities.

3.3. The Effect of Priors

In this section, we explore the effect of prior distributions.

It is important to recognize that, in real-world training environments, pilots rarely perform each task and get assessed by several instructors hundreds of times. Therefore, large datasets like the one in Section 3.1 are uncommon. To simulate a more realistic dataset, we proceed as follows: we keep the same setting of 6 pilots, 4 tasks and 3 instructors, with their parameter values unchanged as in Section 3.1. Considering that pilots undergo recurrent training regularly (usually at least once a year, depending on the regulations of different countries), we treat each training session as a "batch." In every batch, each pilot completes all tasks and is evaluated by just one instructor, with each instructor assessing two pilots. For instance, in batch 1, pilots 1 and 2 are evaluated by instructor 1, pilots 3 and 4 by instructor 2, and pilots 5 and 6 by instructor 3. For this section, we simulate the scores for 15 batches to create the small dataset, detailing the instructor assignments in Table 15. This results in a total of 360 scores.

	batch		
pilot	1-5	6-10	11-15
1	instructor 1	instructor 2	instructor 3
2	instructor 1	instructor 2	instructor 3
3	instructor 2	instructor 3	instructor 1
4	instructor 2	instructor 3	instructor 1
5	instructor 3	instructor 1	instructor 2
6	instructor 3	instructor 1	instructor 2

Table 15 Instructor assignment.

Keeping the prior distributions consistent with Section 3.1, wherein all parameters follow Normal(0,1) distribution independently, we then obtain the estimated parameters from this smaller dataset. A comparative review of the estimated parameters between this dataset and the comprehensive dataset is presented in Tables 16 - 19.

	1st	2nd	3rd	4th	5th	6th
True parameter	-0.80	-0.30	0.00	0.20	0.40	1.00
Pilot ability $\hat{\alpha}_i$ with large dataset	-0.79	-0.27	0.17	0.16	0.19	0.86
Pilot ability $\hat{\alpha}_i$ with small dataset	-0.36	-0.36	0.03	0.16	0.17	0.69

Table 16 Comparison of estimated pilot abilities.

	1st	2nd	3rd	4th
True parameter	0.80	0.00	-0.30	-0.50
Task difficulty $\hat{\gamma}_j$ with large dataset	0.78	0.15	-0.22	-0.33
Task difficulty $\hat{\gamma}_j$ with small dataset	0.61	0.03	-0.15	-0.23

Table 17 Comparison of estimated task difficulties.

	1st	2nd	3rd
True parameter	1.00	0.00	-1.00
Instructor severity $\hat{\delta}_k$ with large dataset	1.22	0.12	-1.04
Instructor severity $\hat{\delta}_k$ with small dataset	1.38	-0.19	-0.82

Table 18 Comparison of estimated instructor severities.

Score	*1*	*2*	*3*	*4*	*5*
True parameter	0	-1.00	-0.50	1.00	1.50
Relative difficulty $\hat{\tau}_l$ with large dataset	0	-1.23	-0.63	0.89	1.33
Relative difficulty $\hat{\tau}_l$ with small dataset	0	-1.59	-0.67	0.84	1.56

Table 19 Comparison of estimated relative difficulties.

As expected, we observe that the parameters estimated from a small dataset are not as accurate as those estimated from a larger one. However, in industry, we usually have some foundational knowledge regarding the parameters under assessment. For example, a pilot with more than 10 years of flight experience is expected to possess greater ability than a recent flight school graduate. This inherent knowledge allows us to establish more informed prior distributions, enhancing the accuracy of the estimations.

Using pilot abilities as an example: assume that based on historical performance metrics, we observe that the higher the pilot label number, the more proficient the pilot. Specifically, the 1st pilot is perceived as the least proficient, while the 6th pilot is the most proficient.

In Section 3.1, we used an independent Normal(0,1) as the prior distribution for all pilots to carry out the analysis. In this Section, we assign independent normal prior distributions for the six pilots, with the prior mean being -0.6, -0.2, 0.0, 0.2, 0.4 and 1.0 respectively. These values more closely align with the assigned parameter values from which we generated the dataset in Section 3.1. As for the variance, we consider two distinct settings. The first setting maintains the variances of these priors at 1 as in Section 3.1, while the second adjusts them to 0.2. The discrepancy between the estimated parameters and the assigned parameter values is given by

$$D = \sum_{i=1}^{I} |\hat{\alpha}_i - \alpha_i| \tag{11}$$

where α_i represents the assigned ability for pilot i and $\hat{\alpha}_i$ denotes the estimated pilot ability for pilot i.

Table 20 presents an overview of the estimated pilot abilities under each test condition as well as the discrepancy. The parameter μ_p is used to denote the improved prior means -0.6, -0.2, 0.0, 0.2, 0.4 and 1. We can see that with the small dataset, specifying the improved prior means without changing the variance results in better estimates for some of the pilot abilities and lowers the discrepancy statistic D, but the first three estimated parameters remain unimproved.

Meanwhile, if a smaller variance can be applied along with the improved prior means, the results yield more accurate estimates, as indicated by the discrepancy statistic D. However, the estimated ability for the 1st pilot at -0.54, aligns more closely with the prior mean than with the true value. This suggests that, with limited data, the estimated outcome is significantly influenced by the prior mean and the variance.

Pilot	*1st*	*2nd*	*3rd*	*4th*	*5th*	*6th*	*D*
True parameter	-0.80	-0.30	0.00	0.20	0.40	1.00	
Pilot ability $\hat{\alpha}_i$ with large dataset. Normal (0, 1).	-0.79	-0.27	0.02	0.16	0.19	0.86	0.45
Pilot ability $\hat{\alpha}_i$ with small dataset. Normal (0, 1).	-0.36	-0.36	0.03	0.16	0.17	0.69	1.11
Pilot ability $\hat{\alpha}_i$ with small dataset. Normal (μ_p, 1).	-0.34	-0.30	0.10	0.26	0.31	0.98	0.73
Pilot ability $\hat{\alpha}_i$ with small dataset. Normal (μ_p, 0.2).	-0.54	-0.23	0.02	0.20	0.37	0.98	0.40

Table 20 Comparison of estimated pilot abilities, with different datasets and prior distributions. The discrepancy statistic D is defined in equation (11).

3.4. Competency Score

In this section, we illustrate the procedure of calculating a competency score for a specific competency, using one pilot as an example, and elaborate on the merits of utilizing the competency score as a better approach.

To recapitulate the methodology for calculating the competency score, the following steps are taken:

1. Estimate parameters using model (7) and Bayesian estimation.
2. Obtain the estimated probabilities for each student pertaining to each task, assuming they are graded by an average instructor by inputting the estimated parameters into model (8).
3. Calculate the projected scores for each pilot on each task utilizing equation (9).
4. For a specific competency of a pilot, calculate the competency score based on the projected scores and the corresponding weights of the competency.

Thus, we first determine the estimated task difficulties, instructor severities and relative difficulties using the same small dataset from Section 3.3. When assigning priors, we proceed as follows: based on prior knowledge, we select prior means to be (0.5, 0, -0.3, -0.6) for task difficulties, (0.8, 0, -0.8) for instructor severities, and (-0.8, -0.4, 1, 1.3) for relative difficulties. The variance is 0.2 for all the prior distributions. The results are presented in Tables 21 – 23.

	1st	*2nd*	*3rd*	*4th*
Task difficulty $\hat{\gamma}_j$	0.47	0.01	-0.33	-0.57

Table 21 Estimated task difficulty based on small dataset.

	1st	*2nd*	*3rd*
Instructor severity $\hat{\delta}_k$	0.83	0.00	-0.86

Table 22 Estimated instructor severity based on small dataset.

Score	*1*	*2*	*3*	*4*	*5*
Relative difficulty $\hat{\tau}_l$	0	-1.03	-0.46	1.00	1.64

Table 23 Estimated relative difficulty based on small dataset.

After acquiring all the estimated parameters, we incorporate the estimated parameters from Tables 20-23 into model (8) to determine the estimated probability distributions for the 1st pilot. As highlighted in Section 2.4, we adopt a value of 0 for instructor severity, assuming the pilots are graded by an instructor with average severity. Table 24 presents the estimated probabilities for the 1st pilot performing each of the four tasks.

Parameters			*Estimated probability*				
1st Pilot	Task	Instructor	1	2	3	4	5
-0.54	*0.47*	*0*	0.11	0.04	0.06	0.27	0.52
-0.54	*0.01*	*0*	0.16	0.03	0.06	0.26	0.49
-0.54	*-0.33*	*0*	0.21	0.03	0.06	0.24	0.46
-0.54	*-0.57*	*0*	0.26	0.03	0.05	0.23	0.43

Table 24 Estimated probabilities for the 1st pilot, with ability $\hat{\alpha}_1 = -0.54$.

Subsequently, utilizing formula (9), we calculate the projected scores for the 1st pilot, and the scores for these four tasks are represented as $S_1 = (3.93, 3.76, 3.59, 3.45)$.

Assuming all four tasks associate with the competency of situation awareness, with respective weights of 10%, 30%, 40%, and 60%, the competency score of situation awareness for the 1st pilot, calculated using formula (10), is:

$$\frac{4.06 * 10\% + 3.88 * 30\% + 3.70 * 40\% + 3.56 * 60\%}{10\% + 30\% + 40\% + 60\%} = 3.70.$$

Utilizing the same methodology, we compute the competency score of situation awareness for the 6th pilot, arriving at a value of 4.24. This score is higher than that of

the 1st pilot, a result consistent with our initial settings where the 6th pilot possesses a greater ability parameter than the 1st.

On the other hand, as mentioned in Chapter 1, the current competency-based grading system evaluates each pilot's competencies based on a subjective evaluation of overall performance by the instructor, rather than deriving from specific task scores. Table 25 presents the raw scores of the fourth batch in the small dataset for both the 1st pilot, assessed by the 1st instructor, and the 6th pilot, assessed by the 3rd instructor.

It is evident that both the 1st and 6th pilots have comparable scores, with the 6th pilot attaining a score of 3 for the 1st task. Given that all the four tasks are associated with competency of situation awareness, it is reasonable to say that these two pilots have the same level of situation awareness. However, utilizing our model, we are able to identify that the 6th pilot is indeed more adept in situation awareness than the 1st pilot. This is achieved by accounting for instructor severity, recognizing that the 1st instructor tends to be more lenient than the 3rd instructor.

Task	*1*	2	3	4
1st pilot *1st instructor*	4	5	4	5
6th pilot *3rd instructor*	3	5	5	5

Table 25 Average scores for the 1st and 6th pilot for each task in the forth batch, graded by 1st and 3rd instructor respectively.

Chapter 4.

Discussion

We have discussed a reparametrized version of the Many-facet Rasch Model with Bayesian estimation and presented the calculation of the competency score, based on certain assumptions and what we believe is realistic in practice.

While humans can account for complexities, there's a limit to the number of factors a person can consider simultaneously and consistently. Compared to the current method, a competency score obtained from our model offers many advantages:

1. **Objectivity**. Relying on task grading data, our model mitigates the influence of emotions, personal bias and other subjective factors that can introduce inconsistencies.

2. **Complexity**. The model can consider a large number of parameters simultaneously and account for intricate relationships among those parameters, thereby providing a more nuanced score.

3. **Reproducibility**. This approach allows for examination and replication by fellow researchers, guaranteeing not just reproducibility, but also creating opportunities for further studies.

While the proposed model offers numerous advantages, however, it is not without limitations. As illustrated in Section 3.2, the model is unable to identify all possible probability distributions. However, it should be noted that the test conducted is based on manually selected probabilities. The availability of real data from the industry may potentially illuminate ways to enhance the model's performance.

Another issue is, with Bayesian estimation, large amount of data is required for accurate parameter estimation. Relying solely on scores from a single test is unlikely to yield reliable estimates, especially for pilots. Since real datasets are not available, we use our simulated dataset to hypothesize that pilots would need scores from approximately 15 test batches to provide sufficient information for parameter estimation.

Consequently, new pilots, who haven't undergone as many tests, may not be effectively assessed under this model.

A potential solution within the aviation industry, where tasks typically remain consistent and the members of instructors don’t change much, is to use data from experienced pilots to calibrate the model. This entails forming a dataset with scores from the experienced pilots and estimating parameters to obtain reliable assessments of task difficulties and instructor severities, leading to the formation of stronger prior distributions for them. The calibrated model can then be used to estimate the abilities of new pilots. For the prior distributions of new pilots, a comparative approach might be effective. Specifically, if new pilots achieve similar grades on tasks as experienced pilots did, and are evaluated by instructors of similar severity, we might infer that the new pilot's abilities are akin to those of the experienced pilots.

Additionally, a further challenge arises from the fact that, while pilots routinely perform certain tasks like take-offs and landings in every test, other tasks may not be executed as frequently. This irregularity in task performance can result in insufficient data for accurate estimation of scores for these less frequent tasks. To address this, one potential solution is to categorize tasks, using these categories as substitutes in the estimation process. There are two possible ways of doing this.

1. Grouping by Similar Competencies: Tasks that are similar, particularly in terms of their associated competencies, can be grouped together. This approach assumes that tasks requiring similar skills or knowledge are equivalent for estimation purposes.

2. Grouping by Difficulty Level: Subsequently, tasks might be grouped based on their level of difficulty. This method groups tasks that are perceived to be of similar challenge, assuming they require comparable skill levels from the pilot.

Furthermore, there are several aspects that warrant additional exploration. One such area is the impact of time on task scores. With pilots gaining flight experience, it's plausible that their skills enhance over time, resulting in improved scores for the same tasks. Thus, a possible avenue for future research lies in the time-weighted tasks.

Specifically, more recent task scores might be given higher weight, whereas older scores might be considered less critical in the assessment.

Another aspect warranting further investigation in this project is the identifiability of the estimated parameters. In Section 3.2, when assessing the model's capabilities, employing a small variance, such as 0.2, in the prior distributions results in differing estimates for pilot ability, task difficulty, instructor severity, and relative difficulty. Despite these variations, the resultant estimated probabilities are notably similar, even though none of them precisely align with the manually selected probabilities.

References

Ackerman, T., Ma, Y., Ma, M., Pacico, J. C., Wang, Y., Xu, G., Ye, T., Zhang, J. & Zheng, M. (2023). Item Response Theory. In R. J. Tierney, F. Rizvi & K. Ercikan (Eds.), *International Encyclopedia of Education* (4th Ed., pp. 72-85), Elsevier.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Defalque, H. (2017). *Competency-based training and assessment* [PowerPoint Slides]. ICAO. https://www.icao.int/ESAF/Documents/meetings/2017/LOC-I%20and%20UPRT%202017/Updated%20Documents/Amdt%205%20to%20PANS-TRG%20v2.pdf

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. New York: Peter Lang Pub. Inc.

ICAO. (2018). *Competency Based Training* [PowerPoint Slides]. ICAO. https://www.icao.int/MID/Documents/2018/CBT%20ATCO%20and%20ATSEP%20Wksp/Module%203%20-%20Intro%20to%20CBT.pdf

Linacre, J. (1989). *Many-faceted Rasch measurement*. San Diego: MESA Press.

Lord, F. M. (1950). A method for estimating from speeded test data the power condition scores and item difficulties. *ETS Research Bulletin Series*, i-7.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, 7, x, 84.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517-548.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. Springer, Berlin, pp 153–164.

Rasch G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, 4, 321-333.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34, 1-97.

Uto, M. & Ueno, M. (2020). A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, 47, 469–496.

Uto, M. (2022). A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01997-z