IJOESS *Year: 8, Vol:8, Issue: 29* **SEPTEMBER 2017**

EVALUATION OF SAFETY PERFORMANCE INDICATORS OF FLIGHT TRAINING ORGANIZATIONS IN TURKEY

Ender GEREDE
Doç. Dr., Anadolu Üniversitesi, egerede@anadolu.edu.tr,
ORCID: 0000-0002-8211-8875

Mehmet YAŞAR
Araş. Gör., Anadolu Üniversitesi, mehmet_yasar@anadolu.edu.tr,
ORCID: 0000-0001-7237-4069

Received: 22.04.2017 *Accepted: 18.07.2017*

ABSTRACT

Safety is a precondition for taking full advantage of air transportation. The International Civil Aviation Organization (ICAO) has now adopted performance-based approach in addition to the regulatory-based with a view to improving safety and decreasing the number of accidents. This new approach requires the implementation of the Safety Management System (SMS), which focuses on increasing safety performance under real-life conditions. The success of its implementation depends on measuring safety performance accurately, and the accurate measurement relies on designing the Safety Performance Indicators (SPI) appropriately. The ultimate aim of this study is to enable design of more effective indicators, and thus to contribute to the improvement of Safety Management Systems. The scope of this study is limited to flight training organizations (FTOs) in Turkey. For the purpose of this study, we first produced a list of SPIs used to measure safety in FTOs, on the validity and significance of which experts were asked to come to an agreement. We then adopted an academic approach to the data collected to determine whether the list of SPIs is valid enough to measure safety under real-life conditions. The Delphi Method was used for data collection and analysis in this study. At the end of five Delphi round extending over a period of one year, we obtained 64 SPIs, which – participants agreed that – measure safety accurately. The research has shown that there were some mistakes and problems related to SPI design, which probably derives from the fact that SPI design is a new quite new requirement in aviation safety. The results of this study are expected to contribute to the improvement of SPI design and hence the success of SMS.

Keywords: Safety Performance Indicators, Performance-Based Approach, Delphi Method, Aviation Safety Management System, Flight Training Organizations.

Gerede, E. ve Yaşar, M. (2017). Evaluation of Safety Performance Indicators of Flight Training Organizations In Turkey, International Journal Of Eurasia Social Sciences, Vol: 8, Issue: 29, pp. (1174-1207).

EMNİYET PERFORMANS GÖSTERGELERİNİN DEĞERLENDİRİLMESİ: TÜRKIYE'DEKİ UÇUŞ EĞİTİM OKULLARI BAĞLAMINDA BIR İNCELEME

ÖZ

Havayolu taşımacılığının sunduğu faydaların elde edilmesi havayolu taşımacılığının emniyetli bir biçimde sürdürülebilmesi ile mümkündür. Uluslararası Sivil Havacılık Örgütü emniyeti daha da artırmak ve kaza sayılarını azaltabilmek amacıyla emniyet yönetiminde dünyayı olması gerektiği gibi gören paradigmanın ürünü olan düzenleme temelli yaklaşımın yanına dünyayı olduğu gibi gören paradigmanın ürünü olan performans temelli yaklaşımı eklemiştir. Bu yaklaşım gerçek hayat şartlarında emniyet performansını artırmaya odaklanan Emniyet Yönetim Sisteminin (EYS) uygulanmasını gerektirmektedir. Bu kapsamda emniyet performansının doğru bir biçimde ölçülmesi son derece önemlidir. Doğru bir ölçüm ise Safety Performance Indicators'ların (SPIs) geçerli bir biçimde tasarlanmasına bağlıdır. Bu çalışmanın kapsamını Türkiye'deki uçuş eğitim organizasyonları -flight training organizations- (FTOs) oluşturmaktadır. Çalışmada FTO'larda emniyeti ölçmek için hangi emniyet performans göstergelerinin kullanıldığı ve çalışmaya katılan uzmanlara göre hangilerinin daha geçerli ve önemli olduğu konusunda üzerinde hemfikir olunacak SPI'ların listesi çıkartılmıştır. Sonrasında elde edilen veriler akademik bir yaklaşımla incelenerek emniyeti ölçecek geçerlilikte olup olmadıkları tespit edilmiştir. Araştırmada verilerin toplanmasında ve analiz edilmesinde Delphi tekniği kullanılmıştır. 1 yıllık zaman dilimini kapsayan 5 Delphi turu sonunda uçuş okullarında emniyeti doğru bir biçimde ölçtüğü düşünülen ve katılımcıların üzerinde hem fikir oldukları 64 SPI elde edilmiştir. Araştırma sonunda SPI tasarımında bazı hata ve sorunların olduğu tespit edilmiş ve bu durum SPI tasarımının yeni bir gereklilik olmasına bağlanmıştır. Çalışmanın SPI tasarımındaki ve dolayısıyla EYS'deki başarıyı artıracağı düşünülmektedir.

Anahtar Kelimeler: Emniyet Ölçüm Göstergeleri, Performans Temelli Yaklaşım, Delphi Tekniği, Havacılıkta Emniyet Yönetim Sistemi, Uçuş Eğitim Okulları

INTRODUCTION

Air transportation not only makes significant contributions to countries and humanity in economic, social and cultural terms but also supports the development of trade and tourism (O'Connor, 2001: 17; IATA, 2007). This is mainly because it ensures rapid entry to markets as a result of spatial connections, as well as providing time advantage (Lohmann et al., 2009: 205; ATAG, 2012). However, in order to take full advantage of air transportation, there is a need to ensure safety. Otherwise, air transportation will lose trust and people will not prefer air transportation as human life is threatened. This, moreover, jeopardizes expected benefits of air transportation as costs of the system will increase. Therefore, it is important to maximize safety in air transportation. It is required to manage safety in order to enhance safety in air transportation. For, air transportation activities undoubtedly bring about safety hazards and risks (Cova and Conger, 2004: 17.1). Safety management relies, above all, on accurate measurement of safety (Gerede, 2016: 160), considering that if you cannot measure it, you cannot manage it (Emil et al., 2005: 9). Safety measurement requires accurate and valid safety measurement indicators. When indicators are problematic, measurements are inaccurate. Inaccurate measurement results certainly invalidate the function of measurement.

What has put greater emphasis on the design of accurate and valid safety performance measurement indicators is the transition to performance-based approach in safety management. The Safety Management System (SMS), the product of transition to the new approach, requires measuring safety by means of safety performance indicators, as well as monitoring and enhancing safety performance. In other words, the success of SMS depends on the accuracy and validity of indicators used for safety performance measurement. The success of SMS will obviously impinge upon aviation safety.

It is considered that SMS is a quite new tool in safety management, and thus, there is currently limited information and experience with regard to safety performance indicators (SPIs). This study seeks to answer the following questions, taking flight training organizations (FTOs) as a case in point:

- What are the safety performance measurement indicators used to measure safety in Turkish FTOs?
- Which SPIs, do safety experts in FTOs believe, are better at safety measurement?
- Which of these indicators are believed to be of greater priority in safety measurement?

I. PERFORMANCE-BASED APPROACH AND SAFETY PERFORMANCE MEASUREMENT

A. Performance-Based Safety Management Approach

The design of accurate and valid safety performance measurement indicators has become more necessary and important with the paradigm shift in safety management. The traditional safety management paradigm sees the world as it ought to be. It defines the ideal world order by means of regulations intricately and prospectively, and requires aviation organizations to keep up to expectations of the ideal world order. The guarantee of achieving the ideal world order relies on the principle of “compliance with regulations”. In this approach, known as the regulatory compliance-based approach, it is believed that safety enhances when compliance with regulations is ensured, audits are performed to assure compliance with regulations, and non-compliance is penalized (ICAO, 2013a: 2-32; ICAO, 2009: 3-10, 3-11, 3-13; Gerede, 2015a; Gerede, 2015b).

Nevertheless, these expectations are not tailored to real-life conditions. For, this approach is predicated on auditing of existing documentation, rather than examination of processes and real safety performance. It is not realistic to believe that the aviation system which operates in a technical and social environment, functions properly without any distortion from predetermined targets. What is more important than compliance with regulations is to improve safety performance over time under real-life conditions (Gerede, 2016: 161).

The new safety management paradigm does not see the world as it ought to be, but as how it really is (ICAO, 2009: 3-13; ICAO, 2013a: 2-5, 2-32). This shift in paradigm brings about a change in our perspective of organizations and human beings. The new approach accepts the reality that human beings and organizations can make errors and even violate rules. It thus intends to detect safety hazards and predict safety risks to take measures before they happen. It focuses on processes proactively rather than on the black-and-white dualism adopted to ensure compliance with regulations. The new approach further suggests that there is a need to concentrate on root causes of errors and violations proactively rather than take a punitive action based on audits. What is essentially expected from aviation organizations is to enhance safety performance continuously as well as complying with regulations. In a nutshell, the new safety management approach goes beyond compliance with regulations and puts specific emphasis on performance-based approach (ICAO, 2013a: 2-5, 2-32). This approach measures safety performance by means of appropriate indicators developed for different fields of activity and processes, monitors it over time, and makes use of proactive tools to achieve predetermined goals. The main objective is to enhance safety performance and ensure compliance with regulations as well. To achieve this, there is a need to design accurate and valid safety performance measurement indicators. Otherwise, measurement efforts will be unsuccessful, and inaccurate measurement results will have negative impacts on SMS performance.

B. Measurement of Safety Performance

Safety performance measurement is determining the degree of safety in the activities of an aviation organization based on safety indicators that take predetermined values as a reference (ICAO, 2013b: 1-2). SPIs are the variables that provide data about the degree of safety performance based on certain qualitative or quantitative values. OECD defines SPIs as tools that indicate how the level of safety has changed over time (OECD, 2008). ICAO defines SPIs as parameters used to measure safety performance (ICAO, 2013a).

The primary aim of safety performance measurement is the controlling function of management. One of the significant functions of management is to assess to which extent goals are achieved and to produce feedback (Daft, 2008). In order to do this, there is a need to measure safety performance by means of accurate and valid indicators. Organizations need to know what happened in the past, what has been happening currently, and what is likely to happen in the future so as to achieve their goals. This is the only way to develop understanding of what needs to be done and which resources need to be used in order to decide what to do (Hollnagel and Woods, 2006).

As explained above, the change in safety management approaches has entailed the measurement of safety performance. Civil aviation authorities will provide safety goals to aviation organizations and ask them to achieve these goals within the frame of SMS (Gerede, 2015a; Gerede, 2015b). Furthermore, in order to protect public health, regulatory and auditing authorities operating on behalf of the state need information on to which extent safety is ensured and whether safety is at risk in aviation organizations. To this end as well, it is required to measure safety performance (Gerede, 2016: 161).

There is a need to measure past safety performance of an aviation organization to determine the success of past hazard and risk analyses as well as risk mitigation measures and to draw lessons for the future. It is possible to evaluate the success of safety and risk management by measuring whether measures taken worked or not, or were implemented or not. Based on such data, one can make predictions for the future. There is a need to know past and current safety performance and to compare these data with goals and targets in order to decide where, why and to which extent to intervene. By doing so, it is possible to predict and prevent future accidents and severe incidents. Safety performance measurement is also required for generating feedback. Furthermore, information on safety performance of an organization allows employees to enhance awareness and dedication to safety, and hence contributes to the strengthening of positive safety culture in an organization (Safety Management International Collaboration Group, 2013: 3-4; Hale, 2009; Gerede, 2016: 161). In addition, the measurement of past safety performance facilitates future-oriented trend analyses and decision making. Measurement is also a precondition for an aviation organization to benchmark its performance with others' performance (Heavisides and Thompson, 2011). Benchmarking is an effective way of learning from other organizations.

C. Characteristics of Efficient Safety Performance Indicators

Safety performance indicators (SPIs) must be, above all, valid and reliable. Validity refers to the degree of measuring accurately the quantity that needs to be measured. Reliability refers to the extent to which measurements of the same quantity yield consistent results in different contexts (Hale, 2009). In order to ensure validity, there is a need for scientific or practice-based evidence verifying that a performance indicator yields undesirable results (Dyreborg, 2009).

Performance indicators must be sensitive enough to perceive changes that occur in a short period of time (Hale, 2009). The fact that a flight training organization (FTO) has not experienced any fatal accidents in the last one, two or ten months does not mean that it is a safe organization. Based on this data, it is not possible to conclude that the organization will not be involved in any accidents in the near future. Latent hazards that threaten safety in an organization may cause fatal accidents. High safety performance based on the number of past accidents does not provide any information about latent hazards (Gerede, 2016: 164). Other requirements that safety performance indicators must have are presented below (Hale, 2009; Step Change in Safety, 2003; Heavisides and Thompson, 2011; Øien et al., 2011; J.G. Verstraeten et al., 2014; Gerede, 2016: 164):

- The design and naming of indicators must not lead to ambiguities and several interpretations. Indicators need to be easily comprehensible and easily communicated to stakeholders.
- The data required by an indicator must be readily available in the organization and collected or generated easily.
- It is advantageous if indicators can be expressed in terms of quantitative data.
- In order for values provided by indicators to be statistically significant, there is a need for substantial amount of data.
- Indicators must allow benchmarking between different departments of an aviation organization, its stations in different locations, and other organizations.
- Indicators must be sensitive enough to show any increase or decrease in performance when there is a change.

D. Classification of Performance Indicators

Lagging indicator and leading indicator are the most common categories of indicators used in safety performance measurement. Lagging indicators provide information about how often incidents and accidents that need to be prevented occurred in the past. With regard to time scale, lagging indicators are on a position that succeeds harm (Reiman and Pietikäinen, 2012). Among lagging indicators are the numbers of fatal aircraft accidents, deaths and severe injuries. Leading indicators concentrate on harmful incidents that have not happened yet but are likely to happen in the future, rather than on past incidents (Reiman and Pietikäinen, 2012). The UK Health and Safety Executive reports that leading indicators are related to activities carried out directly to improve safety (UK Health and Safety Executive, 2006). The design philosophy of leading indicators is based on the supposition that future safety is threatened when activities carried out to maintain risks at acceptable levels fail or are interrupted. Leading indicators are potentially more valuable in performance-based proactive safety management approaches. The indicators that enable us to determine the quality and quantity of activities performed to maintain safety at the acceptable level are qualified as leading indicators. To exemplify, the rate of controlling safety equipment according to a pre-defined program constitutes a good example of leading indicators (Gerede, 2016: 165).

Lagging indicators provide information on whether past activities are successful in safety enhancement and whether safety goals are achieved, but do not include adequate information on why these activities are effective or not (Hale, 2009). This is one of the negative aspects of lagging indicators. On the other hand, the lack of a strong connection between leading indicators and safety performance is also likely to yield negative results (Verstraeten et al., 2014).

Another type of classification concentrates on the relationship of processes and outcomes with safety performance in organizations. This may be seen as an alternative naming for lagging and leading indicators. The design of outcome indicators is based on counting the number of incidents that occur as a result of production and safety enhancement activities. Measurement based on these indicators provides information on to which extent an organization's safety goals are achieved, but does not include information on the factors underlying the performance. On the other hand, activity indicators, like leading indicators, attempt to measure the success of safety enhancement activities (OECD, 2008). Safety performance measurement with only outcome indicators may be misleading when safety enhancement efforts are unsuccessful. If no incident has occurred in the organization yet, outcome indicators are likely to show that safety performance is high at that moment. However, if safety enhancement efforts are unsuccessful, incidents that compromise safety are likely to occur in the future. Outcome indicators focus on aviation organization's outcomes, and activity indicators focus on inputs and processes (Erikson, 2009).

The main characteristic of performance-based safety management approach is that it is proactive rather than reactive. Proactivity requires prediction of the future, and hence requires indicators that provide hints about how safety performance will progress in the future. Indicators that bear these characteristics are known as proactive indicators and have common characteristics with leading indicators (Verstraeten et al., 2014). Measurements with proactive or leading indicators contribute to the generation of cautionary information regarding possible incidents. On the other hand, reactive indicators are more analogous to lagging indicators (Verstraeten et al., 2014). They are used to perform reactive measurements directed to the past after the system produces an unsafe outcome and this becomes obvious.

SMS, which the ICAO obliged some aviation organizations to adopt, uses a different classification of indicators. SMS classifies indicators by the degree of significance of any negative events: high-consequence indicators and lower-consequence indicators. It may be said that lower-consequence indicators fulfill the function of leading indicators. Forgetting to lock the door when leaving home or opening up flaps in landing is the result of a lapse of memory. However, the two situations are likely to lead to different consequences. Thus, counting the number of lower-consequence errors and violations induces a proactive indicator. Furthermore, given the fact that the number of lower-consequence occurring is greater than the number of high-consequence occurring (for instance the number of fatal accidents), it is possible to conduct trend analyses and successful predictions.

Lower-consequence indicators and leading indicators are analogous whereas high-consequence indicators and lagging indicators are analogous.

E. Relationship between Amount of Output and Safety Performance

If the rate of exposure augments, the possibility of unsafe outcomes, in other words risk, is likely to increase – even when hazards in the system remain stable. For example, if an FTO does not operate any flights, the possibility of unsafe event related to flight operation is reduced to zero, and hence, there are no risks. The saying "the safest flight is the one that does not take place" has been used to refer to such a link. Risk is affected by the amount of production (ICAO, 2013a: 2-13). As production increases in a field of activity, risk also increases. To ensure benchmarking in safety management, the value of safety performance must be calculated when the amount of production is at a certain level. That is why it is needed to add the production parameter measuring the relevant operation volume to safety performance indicators.

III. METHODOLOGY

A. Data Collecting and Analysis

In this study, the Delphi method was used to collect data and conduct consensus analysis. The Delphi is a method used for data collecting to investigate a broad and complex problem when the researcher needs to aggregate opinions from a group of individuals with different experience and expertise that does not have a history of adequate communication and when time and cost factors do not allow frequent group meetings (Linstone and Turoff, 2002). This method enables experts to reach a compromise on a topic without coming together in real time and place and without affecting each other (Hung et al., 2008: 192; Şahin, 2001: 216). The Delphi is a method designed to yield an expert opinion in a systematic way (Şahin, 2010: 443).

The Delphi method is based on a series of feedback-providing questionnaires to gather expert opinion on a topic, analyze these opinions, and use them to generate new ideas (Powell, 2003: 377). Owing to feedback, participants have the chance to reconsider their replies (Hsu et al., 2007: 1-2). Questionnaire design focuses on opportunities, solutions and predictions. Each questionnaire builds upon the results of the former questionnaire. The process is completed when a consensus is achieved, adequate information is collected, and theoretical satisfaction is attained (Skumolski et al., 2007: 2; Löfmark and Mårtensson, 2017: 83).

The Delphi is particularly preferred when the researcher is likely to encounter one of the situations defined below (Linstone and Turoff, 2002: 4):

- The problem cannot be solved by analytical techniques, and there is a need for subjective evaluations made by experts.
- It is not feasible for participants to come together to generate solutions for a problem.

- Factors such as time and cost do not allow group meetings.
- There is a need for a heterogeneous group of participants to ensure the validity of research results, and to avoid that some individuals have dominance over others in terms of personality and number.

All above factors played a role in selecting the Delphi method for this study. The principle of confidentiality was observed in order to foster validity of the research and comply with ethical principles. Participants' identity, as well as anonymity, is protected. The Delphi method allowed experts to reach a consensus without coming together and without influencing each other. The fact that opinions of experts were not affected by others' opinions enhances validity of the research.

In the first place, the experts whose opinions would be sought throughout the study (the Delphi panel) were selected. Then, the link of an internet-based questionnaire was sent to participants by e-mail. Participants were first asked which safety performance indicators they used in the FTO and asked to write each indicator one under the other in the form. In the second place, the replies of all participants were listed to obtain a set of data.

Subsequently, the SPI list was resent to participants who were asked to mention which ones were safety performance indicators. Consensus analyses were conducted with the data, as a result of which the indicators on which participants did not agree were excluded from the list in a series of Delphi rounds. The last round collected data on which indicators in the final list were more important than others for successful measurement of safety.

B. Selection of the Expert Panel

The Delphi participants are selected based on their expertise regarding a subject matter (Hatcher and Colton, 2007: 573). Given that expert opinion is required, purposive sampling that allows researchers to select participants qualified enough to answer research questions is preferred over random sampling that represents the population (Skumolski et al., 2007). In this study, the Delphi panel comprises experts employed in FTOs in Turkey. All FTOs licensed by the Directorate General of Civil Aviation (DGCA) were included in this study through complete count. The up-to-date list of licensed FTOs, contact information and names of managers responsible for setting up and managing the SMS were obtained from the DGCA. The DGCA was also asked to encourage FTOs and experts in these organizations to participate in the study. Data on which indicators were used or prioritized to measure safety performance in FTOs were gathered from experts employed in these organizations selected through purposive sampling, considering that they had thorough knowledge of SMS and safety performance measurement. In this respect, with a view to enhancing validity, reliability and plausibility of the study, participants were selected from amongst managers in SMS, quality assurance and flight operation departments of FTOs in consideration of purposive sampling principles. Twenty participants included in the

Delphi panel potentially had considerable knowledge of safety measurement indicators, given their position and responsibilities. A faculty member that previously worked as an SMS manager and is now specialized in safety management was also included in the study. The Delphi panel initially consisted of 37 participants, but 20 experts completed all Delphi rounds. Thus, the final Delphi panel comprised 20 participants. Data related to participants of the Delphi panel are presented in Table 1.

Table 1. Delphi Panel Demographics

Professional Position				
SMS Manager	**Quality Manager**	**SMS Expert**	**Quality Expert**	**Faculty – Pilot**
5	4	5	2	1 - 3
Educational Background				
High School	**Associate Degree**	**Bachelor's**	**Master's**	**Ph.D.**
-	-	16	1	3
Professional Experience (Years)				
0-5	**6-10**	**11-15**	**16-20**	**+21**
5	7	2	3	3
Experience in Aviation Safety (Years)				
0-5	**6-10**	**11-15**	**16-20**	**+21**
9	10	-	1	-

Table 1 shows that 25% of the Delphi panel consisted of SMS managers, 25% SMS experts, 20% quality managers, 10% quality experts, and 20% faculty members and pilot trainers. In the panel, 80% of the experts held a Bachelor's degree. The rate of master's degree holders was 5%, and doctoral degree holders was 15%. Data related to participants' work experience in the aviation industry show that 35% of participants worked for 6-10 years, 25% for 0-5 years, 15% for 16-20 years or over 20 years, and 10% for 11-15 years. The distribution of their experience in aviation safety was as follows: 50% between 6 and 10 years, 45% between 0 and 5 years, and 5% between 16 and 20 years.

C. Consensus Criteria

The indicators on which participants did not come to a consensus were eliminated in sequential Delphi rounds after consensus analysis was conducted on data obtained through a seven-point Likert-type scale. Arithmetic mean, standard deviation and interquartile range (IQR) values were calculated based on participants' scoring from 1 to 7. The results were sent to the experts in the next round in order for participants to make more qualified decisions.

Researchers adopt different approaches to the procedures and criteria of consensus analysis. In a study on near miss reporting system in the maritime industry, Terzi and Gazioğlu (2014) used five-point Likert scale and defined a two-way consensus. These consensus criteria are shown in Table 2.

Table 2. Consensus Criteria Based on Median, IQR and Frequencies

Consensus on approval	Consensus on rejection
If median is ≥4 and IQR is ≤ 1	If median is ≤ 2 and IQR is ≤ 1
If IQR is ≤ 1.5 and 4-5 frequency is ≥ 80%	If IQR is ≤ 1.5 and 1-2 frequency is ≥ 80%

Reference: Terzi and Gazioğlu, 2014: 45.

The table shows that there is a positive consensus when the median is equal to or greater than 4 and the interquartile range is equal to or less than 1. If these values are not obtained and the IQR is greater than 1, then the researchers have a look at the second consensus criterion. A positive consensus is achieved when the IQR is equal to or less than 1.5, and at the same time, the frequency of 4-5 values counts for 80% of all answers.

The research also seeks the “negative” way of consensus, in other words lack of consensus. There is a consensus that there is no consensus when the median is equal to or less than 2 and the IQR is equal to 1 or less than 1. If these values are not obtained, the researchers have a look at the second criterion. A negative consensus is achieved when the IQR is equal to or less than 1.5, and the frequency of 1-2 values counts for 80% of all answers.

Nevertheless, some studies seek full consensus. In a study conducted at Ege University (Saçaklıoğlu et al., 2005), IQR was used as a criterion, and the degrees of consensus were defined as follows in Table 3:

Table 3. Criteria and Degrees of Consensus Based on IQR

IQR	Degree of Consensus
0.0	Full
>0.0-0.5	Very Strong
>0.5-1.0	Strong
>1.0-1.5	Moderate
>1.5	Weak

Reference: Saçaklıoğlu et al., 2005: 94.

In another study, Sharkey and Sharples (2001) took standard deviation as a consensus criterion. In this respect, they report that there is high consensus on items that are less than one standard deviation away from the mean value, medium consensus on items that are between one and two SDs away from the mean value, and low consensus on items that are more than two SDs away from the mean value. Gençtürk and Akbaş (2013: 346) use five-point Likert scale and defined consensus criteria based on the IQR. They accept that there is no consensus on items whose IQR is below 1.2. Linz (2012: 30) also uses five-point Likert-type scale in a research conducted to identify 2025 scenarios in the aviation sector, and reports that consensus is reached on items with an IQR equal to or less than 30%.

Among studies in which seven-point Likert-type scale is used, Şahin's (2010) study related to educational sciences plays a guiding role. In that study, the criteria are based on medians, IQRs and frequencies, as defined below in Table 4.

Table 4. Consensus Criteria in Studies Conducted with 7-Point Likert Scale

Presence of Consensus	Median	IQR	Frequency
Yes	≥ 5	≤ 1.5	-
	≥ 5	≤ 2.5	≥ 70% (5-7)
No	≤ 3	≤ 1.5	-
	≤ 3	≤ 2.5	≥ 70% (1-3)
Neutral	= 4	≤ 2.5	-

Reference: Şahin, 2010: 446.

In this study, seven-point Likert-type scale is preferred to determine the indicators that are considered to measure safety performance accurately. Seven-point Likert-type scale is chosen considering that the expert panel is equipped with knowledge required to make fine-tuning evaluations about aviation safety. Furthermore, in Likert-type scales, the use of greater number of response choices allows participants to make a choice among more response opportunities, and hence intends to improve internal consistency (Köklü, 1995: 90). For the purpose of this study, interquartile ranges, medians and frequencies are used together as consensus criteria. Table 5 provides ranges regarding consensus criteria used in analysis.

Table 5. Consensus Criteria Used in the Research

	Median	IQR	Frequency	Standard Deviation	Arithmetic Mean
1	5	≤ 1.5			
2	5	≤ 2.5	≥ 70% (5-7)		
3				≤ 2	≥ 5

Accordingly, consensus has been reached on an item:

- when the median is equal to or greater than 5 and the IQR is equal to or less than 1.5,
- when the median is equal to or greater than 5 and the IQR is equal to or less than 2.5, and the frequency of 5-7 values counts for at least 70% of all replies,
- when the arithmetic mean is equal to or greater than 5, and the standard deviation is less than 2.

D. Application of the Delphi Study

In the first Delphi round, participants were asked, in an open-ended question, to mention safety performance measurement indicators that they used and/or suggested using in their FTOs with a view to obtaining a pool of items that contain safety performance indicators. Participants were asked to enter at least 10 SPIs in a web interface page designed for the study. The web interface page did not allow participants to skip to the next page without entering at least 10 SPIs. Several SPIs in the list obtained in the first round were related to helicopter flight training. These items were excluded from the list before the second Delphi round. Furthermore, the items that were exactly the same or analogous were reduced to a single indicator item that represents the indicator specifically.

In the second round, the list of indicators obtained in the first round was sent to the panel. Participants were asked to rate the SPIs from 1 (invalid, the least effective,) to 7 (valid, the most effective,). The aim was to obtain expert opinion on which SPIs measure safety performance validly. Consensus analysis were conducted on rating data obtained in the second round. As a result, the SPIs on which participants agreed that it was invalid (ineffective) and disagreed that it was valid (effective) were eliminated from the list.

In the third round, the list of SPIs obtained as a result of eliminations in the second round were resent to participants together with means, standard deviations and medians of scores. Participants were asked to rate the items again, taking account of statistical data. The website provided them with detailed information on how to interpret the statistical data. The third and fourth rounds were repeated in the same way in order to obtain a final list of SPIs on which participants reached a consensus.

In the fifth – and last – round of the study, participants were asked to rank in the order of perceived importance the SPIs obtained at the end of the fourth round. The rating averages in the last round were used to rank the SPIs, where the highest score determined the indicator at the top of the list. Five Delphi rounds extended over 13 months in total. This process yielded a list of 64 SPIs that are believed to measure safety performance accurately and validly.

IV. FINDINGS AND INTERPRETATIONS

A. Findings of the First Delphi Round

In the first Delphi round, after participants replied to demographic questions, they were asked the following open-ended question: "Please indicate the safety performance indicators that you use or suggest using to measure safety performance accurately in the flight training organization where you work." A total of 248 SPIs were obtained at the end of one-month period. In this list, 106 SPIs were eliminated, and 142 SPIs were used in the second round. Table 6 provides 142 SPIs that were filtered from the list obtained in the first round and resent to participants in the second round. SPIs were eliminated from the initial list because either they were related to helicopter flight training or they were irrelevant. The items that were exactly the same or had the same meaning were reduced to a single item that represents the indicator specifically.

B. Findings of the Second Delphi Round

In the second Delphi round, a seven-point Likert-type scale was used to collect data regarding the list of SPIs from participants. The questionnaire included the following instruction: "Please rate each of the following items from 7 (this is an exactly valid safety performance indicator) to 1 (this is not a valid safety performance indicator) in order to assess whether it is a valid safety performance indicator.” Participants rated 142 items in the list during a period of two months. The researchers eliminated 44 SPIs on which participants did not reach a consensus. The remaining 98 SPIs were used in the third round. Table 6 presents data related to SPIs obtained in the second round. Consensus criteria are based on frequency percentage, mean, standard deviation, interquartile range (IQR) and median (Q2). SPIs on which participants did not come to an agreement are marked by an asterisk (*) in Table 6.

Table 6. Data Obtained in the Second Delphi Round

Safety Performance Indicators	Production Amount	Validity	Type of SPI	5-7 Frequency Percentage, Mean and SD			Interquartile Range			
				5-7	SD	Mean	Q1	Q2	Q3	IQR
1- Non-injury accidents/incidents (material damage) (Number/Flight hour)	1	2	1a, 2a, 3a, 4b	86	1,32	6,17	6	7	7	1
2- Evaluation of meteorological conditions in solo flights	0	1-0	1b, 2b, 3b, 4b	81	1,89	5,5	5	6	7	2
3- Use of up-to-date and appropriate documents for aircraft type	0	1-0	1b, 2b, 3b, 4b	67	1,99	5,14	4	6	7	3
4- Number of bird strikes	0	0		67	1,73	5,22	4	6	7	3
5- Runway excursion (Number/landing and takeoff)	1	2	1a, 2a, 3a, 4b	89	1,48	5,97	5	7	7	2
6- Number of air miss	0	2	1a, 2a, 3a, 4b	89	1,12	6,33	6	7	7	1
7*- Number of loss of oil pressure	0	0		50	1,85	4,89	4	4,5	7	3
8- Radio communication failure	0	0		61	1,71	5	4	5	7	3
9- Use of wrong taxiway	0	1	1a, 2a, 3a, 4b	72	1,61	5,53	4	6	7	3
10*- Test of fire extinguishing system	0	0		44	2,14	4,06	2	4	6	4
11- Wing strikes	0	1	1a, 2a, 3a, 4b	78	1,75	5,58	5	6	7	2
12- Non-compliance with ATC instructions (Number/Flight hour)	1	1	1b, 2b, 3b, 4b	89	1,45	6,11	6	7	7	1
13- Frequency of aborted flights due to alcohol or psychoactive substance use	0	1	1b, 2b, 3b, 4b	81	1,68	5,72	5	6	7	2
14- Failure to comply with the procedures specified in the	0	1	1b, 2b, 3b, 4b	92	1,47	5,94	5,25	6	7	1,75

Standard Operating Procedures (SOP) on round trips to the training areas and in the traffic pattern										
15- Wing strike during takeoff	0	1	1a, 2a, 3a, 4b	83	1,65	5,81	5	6,5	7	2
16- Number of aircraft collisions with ground vehicles	0	0		83	1,54	5,5	5	6	7	2
17- Critical fuel situation	0	1	1a, 2a, 3a, 4b	81	1,64	5,78	5,25	6	7	1,75
18- Missed touchdown zone	0	1	1b, 2b, 3b, 4b	53	1,72	4,53	4	5	6	2
19- Extended maintenance period due to difficulties in spare part procurement	0	0		28	2,05	3,08	1	2,5	5	4
20- Fatal accidents (Number/Flight hour)	1	2	1a, 2a, 3a, 4a	97	0,96	6,67	7	7	7	0
21*- Closing the first level quality audit findings within the specified time	0	1***0		53	2,09	4,28	2,25	5	6	3,75
22*- Monitoring health of pilots	0	1***0		62	2,05	4,75	3	5	6,75	3,75
23- Failure in taking safety measures during refueling	0	0		78	1,93	5,61	5,25	6	7	1,75
24- Missing the centerline track during takeoff and landing	0	1	1a, 2a, 3a, 4b	64	1,84	5,08	4	5	7	3
25- Non-compliance with operational limits and limits specified in the Pilot's Operating Handbook during the flight	0	1	1b, 2b, 3b, 4b	89	1,28	6,19	6	7	7	1
26- Accidents/incidents involving injuries (Number/Flight hour)	1	2	1a, 2a, 3a, 4a	92	1,17	6,33	6	7	7	1
27*- Number of hazards detected through brainstorming	0	0		50	2,06	4,14	2,25	4,5	6	3,75
28*- Frequency of flight cancellation due to aircraft failure	0	0		53	1,99	4,28	3	5	6	3
29- Damage to aircraft structure during landing	0	1	1a, 2a, 3a, 4b	83	1,37	5,72	5	6	7	2
30- Number of aborts due to FODs on PAT areas.	0	0		64	1,9	5,17	4	6	7	3
31*- Number of GPWS failures	0	0		47	1,98	4,42	3	4	6	3

32*- Number of FOD occurring due to technicians or related staff	0	0		58	2,1	4,86	4	5	7	3
33- Non-compliance with tower instructions	0	1	1b, 2b, 3b, 4b	89	1,56	6,03	5	7	7	2
34- Hit by aircraft propeller (Staff)	0	2	1a, 2a, 3a, 4a	83	1,68	5,92	5	7	7	2
35- Fainting during flight	0	1	1a, 2a, 3a, 4b	72	1,78	5,5	4	6	7	3
36- Presence of unauthorized vehicles, animals or materials on taxiway	0	0		78	1,91	5,31	5	6	7	2
37- Aircraft tire blowout (Number/Flight hour/Takeoff and landing)	0	2	1a, 2a, 3a, 4b	81	1,28	5,83	5	6	7	2
38- Use of improper navigational aids during flight	0	1	1b, 2b, 3b, 4b	75	1,95	5,42	4,25	6	7	2,75
39- Lack of debriefing after the flight	0	1	1b, 2b, 3b, 4b	58	1,74	4,86	4	5	6,75	2,75
40- Pilot allowed to fly at night without taking night flight training	0	1	1b, 2b, 3b, 4b	75	2,23	5,58	4,25	7	7	2,75
41*- Lack of morning briefing	0	1***		64	1,84	4,86	3	6	6	3
42- Aircraft landing gear failure (Number/landing) during landing	1	0		86	1,73	5,83	5	6,5	7	2
43*- Number of days on which flights cannot be operated due to scheduled maintenance	0	0		28	2,2	2,97	1	2	5	4
44*- Documents that must be present in the aircraft are not up-to-date	0	1***		61	2,2	4,58	2,25	5	7	4,75
45*- Frequency of cancelling flights due to health problems	0	1***		42	2,11	3,78	2	4	6	4
46- Tail strikes	0	1	1a, 2a, 3a, 4b	86	1,32	6,03	6	6	7	1
47*- Flameout	0	0		58	2,09	4,89	4	5	7	3
48*- Static grounding in hangars	0	0		56	2,55	4,22	1	5	7	6
49*- Operations of ground support equipment	0	0		44	2,34	3,72	1	4	6	5
50- Engine failure (Number/Flight hour)	1	0		81	1,95	5,92	6	7	7	1
51- Aborted takeoff at high speeds	1	1	1a, 2a, 3a, 4b	78	1,9	5,44	5	6	7	2

(Number/Flight hour)										
52*- Number of fuel spillage during refueling	0	0		67	2,2	4,89	2,5	6	7	4,5
53- Examination of meteorological conditions and NOTAMs in airdromes where flights are scheduled	0	1-0	1b, 2b, 3b, 4b	67	2,25	5,11	3,25	6	7	3,75
54- Number of unauthorized ground traffic activities on the taxiway during flight training	0	0		69	1,87	5,36	4	6	7	3
55- Overspeed	0	0		81	1,86	5,19	5	6	6	1
56- SEP flights operated when meteorological conditions do not meet VFR or VMC requirements	0	1	1b, 2b, 3b, 4b	86	1,69	5,86	5	7	7	2
57- Emergency declaration (Number/Flight hour)	1	0		69	1,59	5,5	4	6	7	3
58- Not checking around in ground movements and exceeding the taxi speed limit	0	1	1b, 2b, 3b, 4b	72	1,66	5,61	4	6	7	3
59*- Number of accidents involving fire fighting vehicles	0	0		39	2,5	3,83	1	4	7	6
60- Contact of aircraft’s wing with the ground during landing	0	1	1a, 2a, 3a, 4b	83	1,51	5,69	5	6	7	2
61- Loss of positional orientation	0	1	1a, 2a, 3a, 4b	81	1,69	5,61	5	6	7	2
62*- Doing nothing related to topics not covered by training manuals	0	1***0		53	2,17	4,42	3,25	5	6	2,75
63- Emergency landing on terrain (Number/Flight hour)	1	2	1a, 2a, 3a, 4b	81	1,79	5,75	5	7	7	2
64*- Frequency of changes in flight program	0	1***		47	1,88	4,22	2,25	4	6	3,75
65*- Not taking into account the critical areas of airplane during the drenching students with water after their first soloing.	0	0		39	2,17	3,61	1	3,5	5	4
66- Being higher or lower than glide path during approach	0	1	1a, 2a, 3a, 4b	64	1,8	4,67	3,25	5	6	2,75

67*- Number of risks at the unacceptable level	0	1***		64	2,35	4,97	3	6	7	4
68*- Aircraft hitting a dog	0	0		58	2,19	4,72	2,5	5,5	7	4,5
69*- Landing in crosswind and heavy rain	0	2***		61	2,1	4,92	4	6	7	3
70- Lack of VFR flight conditions	0	1	1a, 2a, 3a, 4b	75	1,76	5,56	4,25	6	7	2,75
71- Emergency landing	0	2	1a, 2a, 3a, 4b	78	1,85	5,69	5	6,5	7	2
72- Hard landing (Number/Flight hour/landing and takeoff)	1	2	1a, 2a, 3a, 4b	92	1,23	5,97	5	6	7	2
73- Non-compliance with normal / emergency checklist procedures	0	1	1b, 2b, 3b, 4b	89	1,56	6,03	6	7	7	1
74*- Periodic control of fire extinguishers	0	0		64	2,05	4,81	4	5	6,75	2,75
75- Propeller strikes	0	1	1a, 2a, 3a, 4a	89	1,28	6,11	6	6,5	7	1
76- Accidents/incidents involving severe injuries (Number/Flight hour)	1	2	1a, 2a, 3a, 4a	89	1,24	6,33	6	7	7	1
77- Number of air traffic occurring reporting	0	1	1a, 2a, 3a, 4b	81	1,25	5,75	5	6	7	2
78- Exceeding the VFR minima (e.g. flying into clouds, getting too close to the ground) during flight training (e.g. turns, stalls, slow flight)	0	1	1b, 2b, 3b, 4b	92	1,33	6,06	5	7	7	2
79- Number of runway excursions	0	2	1a, 2a, 3a, 4b	92	1,24	6,19	5	7	7	2
80- Missed approach except training purpose	0	2	1b, 2b, 3b, 4b	58	1,83	4,81	3,25	5	6,75	3,5
81- Frequency of cancelling flights due to insufficient knowledge level	0	1	1b, 2b, 3b, 4b	67	2,07	5,06	3,25	6	7	3,75
82- Mid-air collision (Number/Flight hour)	1	2	1a, 2a, 3a, 4a	89	1,24	6,31	6	7	7	1
83- Lack of flight training phases	0	1	1b, 2b, 3b, 4b	69	1,98	5,28	4	6	7	3
84*- Keeping the hangar safe	0	0		56	2,41	4,08	1,25	5	6,75	5,5
85- Pre-flight control	0	1-0	1b, 2b, 3b, 4b	72	2,28	5,14	3,25	6	7	3,75
86- Fire and smoke in the aircraft	0	0		81	1,81	5,78	5	7	7	2
87*- Night flight	0	0		42	2,34	3,53	1	3,5	5	4

88- Violation of NOTAMs/rules (Number/Flight hour)	1	1	1b, 2b, 3b, 4b	86	1,34	6,17	5,25	7	7	1,75
89*- Post-flight control	0	1***0		64	2,22	4,75	3	5	7	4
90*- Exceeding flight time limitations	0	1***		58	2,12	4,58	3	5	7	4
91*- Monitoring the validity of pilot's licenses and training	0	1***0		67	2,14	4,81	3	5	7	4
92- Operating a flight on the first day after return from annual leave	0	1***		39	2,03	3,61	1,25	4	5	3,75
93- Bird strike (Number/Flight hour)	1	0		78	1,63	5,5	5	6	7	2
94*- Closing the first level quality audit findings within the specified time	0	1***0		56	2,13	4,14	2	5	6	4
95- Not considering the crosswind in landing	0	1	1b, 2b, 3b, 4b	75	1,98	5,39	4,25	6	7	2,75
96- Collision on PAT areas (Number/sortie)	1	2	1a, 2a, 3a, 4a	83	1,74	6,06	5,25	7	7	1,75
97- Violation of NOTAMs	0	1	1a, 2a, 3a, 4b	86	1,3	6,11	6	7	7	1
98- Number of stall	0	1	1a, 2a, 3a, 4b	78	1,74	5,69	5	6	7	2
99*- Failures of flight instruments	0	0		56	2,22	4,56	2	5	7	5
100- Missing the center line	0	1	1a, 2a, 3a, 4b	81	1,42	5,86	5	6	7	2
101- Failure of flight control systems (Number/Flight hour)	1	0		86	1,8	5,89	5	7	7	2
102- Flights exceeding MTO limits	0	1	1b, 2b, 3b, 4b	86	1,26	6,31	6	7	7	1
103- Ground accidents	0	2	1a, 2a, 3a, 4a	83	1,42	5,78	5	6	7	2
104*- Number of days on which flights cannot be operated due to breakdown or repair	1	0		28	1,9	2,92	1	2,5	5	4
105*- Broken tow bars	0	0		44	2,14	4,08	2	4	6	4
106- Missed approach (Number/Flight hour/Landing and takeoff)	1	2	1a, 2a, 3a, 4b	56	1,98	4,47	3	5	6	3
107- Reporting at the wrong altitude and point during flight from leaving the training area to landing	0	1	1b, 2b, 3b, 4b	75	1,7	5,53	4,25	6	7	2,75
108- Failure in flight control systems	0	0		78	2,07	5,64	5	7	7	2

109- Collision of aircraft with another equipment (e.g. APU)	0	0		69	1,93	5,36	4	6	7	3
110- Fire and smoke (Number/Flight hour)	1	0		72	2,06	5,53	4	7	7	3
111*- Validity of pilots' medical certificates	0	1***0		67	1,83	4,97	4	5	6,75	2,75
112*- Presence of an insurance policy for hangar	0	0		50	2,38	3,89	1	4,5	6	5
113*- Flight interruption due to long term unfavorable weather conditions	0	0		33	2,04	3,31	1	3,5	5	4
114*- Number of repeating failures (maintenance)	0	0		61	2,35	4,69	2,5	5	7	4,5
115*- Spilling of fuel during refueling	0	0		64	2,19	4,69	3	6	6	3
116*- Solo flight procedures	0	0		53	2,5	4,08	1	5	6	5
117- FOD (living/non-living) in the PAT area (Number/Flight hour)	1	0		72	1,95	5,28	4	6	7	3
118*- Frequency of flight cancellation due to aerodrome facilities	0	0		53	2,15	3,86	2	5	5,75	3,75
119*- Dangerous areas of airplanes	0	0		53	2,39	4,25	2	5	6,75	4,75
120- Number of flaming incidents during refueling	0	0		72	2,16	5,39	4	6,5	7	3
121- Frequency of FOD reporting on the flight line	0	0		67	2,06	5,14	4	6	7	3
122*- Aircraft insurance	0	1***0		42	2,29	3,81	1	4	6	5
123- Number of reported hazards, unsafe events	0	0		75	1,96	5,25	4,25	6	7	2,75
124- Tire blowout	0	2	1a, 2a, 3a, 4b	78	1,5	5,83	5	6,5	7	2
125- Hard landing	0	1	1a, 2a, 3a, 4b	83	1,49	5,81	5,25	6	7	1,75
126- Critical fuel situation (Number/Flight hour)	1	1	1a, 2a, 3a, 4b	78	1,81	5,58	5	6	7	2
127*- Number of hazards whose risk level is lowered	0	0		61	2,02	4,5	2,25	5	6	3,75
128- Misunderstanding of ATC instructions	0	1	1a, 2a, 3a, 4b	81	1,57	5,67	5	6	7	2
129- Incapacity (Number/Flight hour)	1	1	1a, 2a, 3a, 4b	83	1,75	5,81	5,25	6,5	7	1,75
130- Violation of	0	1	1b, 2b,	81	1,62	5,86	5	7	7	2

altitude restrictions			3b, 4b							
131- System failures that affect flight safety (Number/Flight hour)	1	0		81	1,93	5,61	5	6,5	7	2
132- Landing on the wrong runway (Number/Flight hour/Landing and takeoff)	1	2	1a, 2a, 3a, 4b	86	1,35	6,22	6	7	7	1
133- Unauthorized engine start	0	1	1b, 2b, 3b, 4b	72	1,87	5,22	4	6	7	3
134- Getting lost during flight (Number/Flight hour)	1	1	1a, 2a, 3a, 4b	89	1,37	6,11	6	7	7	1
135- Violation of control or training area restrictions (Number/Flight hour)	1	1	1b, 2b, 3b, 4b	80	1,72	5,89	6	7	7	1
136- Runway incursion	0	1	1a, 2a, 3a, 4b	83	1,51	6,11	6	7	7	1
137- Landing gear failure	0	0		78	1,83	5,75	5	7	7	2
138- Exceeding the pilot and aircraft related limitations in solo flight	0	1	1b, 2b, 3b, 4b	83	1,6	6	6	7	7	1
139- Number of incidents	0	2	1a, 2a, 3a, 4b	75	1,81	5,42	4,25	6	7	2,75
140- Engine failure	0	0		83	1,68	6,03	6	7	7	1
141- Non-compliance with the VFR approach and landing restrictions	0	1	1b, 2b, 3b, 4b	75	2,05	5,47	4,25	7	7	2,75
142- Number of undershoot	0	1	1b, 2b, 3b, 4b	69	1,86	5,39	4	6	7	3

In Table 6, the first column shows the SPI suggested by participants, and the second column provides information on whether the production amount is considered in the formulation of SPI. If the SPI formulation includes the production amount, it takes value 1. Otherwise, it takes value 0. The third column shows whether the SPI is valid or not. The researchers examined each item to decide whether it measures safety performance in flight training activities of FTOs, and came to a decision on the validity of each item through negotiation and persuasion. If the SPI is totally valid with regard to the field of activity and processes, it is assigned value 1. If the SPI is clearly invalid, it is assigned value 0.

Some SPIs were problematic in terms of linguistic expression. Yet, it was still possible to understand what they are meant to measure. For instance, SPI 2 which reads as “evaluation of meteorological conditions in solo flights” is obviously related to flight training processes. However, the SPI was formulated in an affirmative way rather than in a negative way. The problem is with the naming of SPI. As explained earlier, an SPI must not lead to any ambiguities. Such SPIs were classified as 1-0.

The results of participants' evaluation of SPIs were assessed with regard to consensus criteria. The researchers eliminated the SPIs on which participants did not reach a consensus. These SPIs are written in bold and marked by '*****' in Table 6. The researchers' evaluation suggests that some of the indicators were eliminated in spite of being valid.

With respect to some SPIs, it is not clear whether the indicator measures safety performance in an FTO or in another field of aviation activity. This is mainly because the field of activity was not indicated clearly. For instance, any unsafe incident suggested by SPI 1, i.e. "Non-injury accidents/incidents (material damage) (Number/Flight hour)", may occur due to a technical failure in the maintenance process. In such a case, SPI 1 is not likely to measure safety performance in flight training, and proves to be invalid. A more accurate and valid SPI statement would be as follows: "An unsafe incident occurring due to pilotage". In Table 6, such SPIs were assigned value 2.

Table 7. Analysis of SPIs in the Second Delphi Round

Categorization of SPIs	SPIs in the Relevant Category		SPIs Not Included in the Relevant Category		Total
	Frequency	Rate (%)	Frequency	Rate (%)	
SPIs taking production amount into account	27	19	115	81	142
SPIs with absolute validity	49	34.5	93	65.5	142
SPIs eliminated although being valid	16	36.36	28	63.64	44
Perfect SPIs	30	41.09	43	58.91	73
SPIs that need to be revised	24	16.9	118	83.1	142
SPIs not eliminated although they need to be	24	24.48	74	75.52	98

Table 7 presents the findings obtained as a result of the analyses of data related to the SPIs collected in the second round. The most interesting finding is that the production amount was almost never taken into consideration in the design of SPIs. In the list, 81% of the SPIs do not include parameters related to production amount. With regard to the field of activity and appropriateness of processes, the rate of SPIs with absolute validity is 34.5%. This is also a noteworthy finding. Given that the remaining SPIs are invalid, safety performance measurement would be inaccurate. Another critical finding is the presence of SPIs that were eliminated because of lack of consensus even though they were valid. Of 44 SPIs eliminated, 16 were valid. Some participants formulated valid and effective SPIs; however, these SPIs were eliminated because of the scores given by the majority of participants.

In the analysis process, the researchers categorized the SPIs that were not eliminated and were considered valid. As explained earlier, the general tendency in safety measurement performance is using lagging reactive indicators that focus on outcomes with high level of importance. The values of such SPIs nevertheless do not provide adequate information on where and why safety performance is low (or high). In such a case, safety performance is measured; however, it is hard to obtain valuable information and performance measurement does not prove to be useful enough. Thus, in the categorization, the SPIs that were leading, process-oriented, proactive and of less importance at the same time were put under the category of perfect SPIs. As is known, SMS is a proactive and predictive safety management tool. In order for SMS to fulfill the expectations, it should be used to predict future successfully based on safety performance data of the past. To do this, SMS needs perfect indicators. In the list, out of 73 items, 30 indicators proved to be perfect. This is a relatively positive finding. Nevertheless, the rate of perfect indicators is quite low in the initial list including 142 SPIs.

The analysis also showed that, owing to the participants' scoring, some SPIs were not eliminated and transferred to the succeeding round even though they were invalid. One would expect that 29 items were eliminated among 98 SPIs that were not eliminated. Furthermore, in the same list, 24 items needed linguistic revision.

Table 8. Types of SPIs Obtained in the Second Delphi Round

Indicator Group	Codes	Type	Number of Indicators	%
1 Lagging Lagging	1a	Lagging	43	58,9
	1b	Leading	30	41.1
		Total	73	100
2 Outcome Activity	2a	Outcome	43	58.9
	2b	Activities	30	41,1
		Total	73	100
3 Reactive Proactive	3a	Reactive	43	58.9
	3b	Proactive	30	41.1
		Total	73	100
4 High Level Low Level	4a	High Level	8	10,95
	4b	Low Level	65	89.05
		Total	73	100

Table 8 shows the types of 73 SPIs that were included in categorization at the end of the second Delphi round. The rate of indicators with lower level of importance in SMS regulations is quite high, i.e. 89%. This is a finding that likely to boost the success of SMS. However, the rate of leading, activity-oriented and proactive indicators is below 50%.

These findings together suggest that participants do not have thorough knowledge of how to design valid and accurate SPIs. The findings point to the following problems in SPI design:

- Participants did not pay attention to production amount,
- Participants took other fields of activity, e.g. maintenance, airport and ground handling, into consideration rather than merely focus on flight training,
- There are ambiguities in SPI statements,
- There is inadequate emphasis on leading, process-oriented and proactive indicators.

C. Findings of the Third Delphi Round

The first two Delphi rounds concentrated on the participants' flawed decisions and problems related to SPI design. From the third round, there was greater focus on obtaining a list of accurate SPIs on which the participants came to an agreement. The following revisions were made in order to obtain a more accurate list of SPIs:

- Dividing into two the SPIs that were logically accurate, but measuring two different magnitudes in one statement. This resulted in an increase in the number of SPIs.
- Combining into one SPI different indicators that were measuring the same magnitude. As a result of these two revisions, the number of SPIs reduced to 88.
- Adding 10 SPIs that were suggested by none of participants, but can potentially be used in safety performance measurement. As a result, the number of SPIs to be sent to the panel increased to 98.
- Adding parameters including amount of production to indicators.

At the end of the third Delphi round, which lasted two months, participants reached a consensus on 66 out of 98 SPIs and eliminated 32 SPIs. Table 9 provides the data obtained after the third round. A total of 32 SPIs, written in bold and marked by '*****' in Table 9, were excluded from the list. Participants eliminated 23 out of 32 SPIs although they were valid. The rate of these items is quite high, and this is a finding that compromises the effectiveness of SPI design. Out of 23 SPIs eliminated, 14 items fell under the category of perfect SPIs (1b, 2b, 3b, 4b categories). This is another finding that jeopardizes the effectiveness of SPI design. In this round, eight out of ten items added to the list by the researchers were eliminated. Two SPIs added to the list by the researchers and transferred to the following round were as follows: "number of self-reports of errors by flight crew per landing and takeoff" and "number of cases where taxiing starts without takeoff authorization per takeoff". After 32 SPIs were eliminated, 68 SPIs were used in the fourth round.

Table 9. Data Obtained in the Third Delphi Round

Safety Performance Indicators (SPIs)	Type of SPI	5-7 Frequency Percentage, Mean and SD			Interquartile Range			
		5-7	Mean	SD	Q1	Q2	Q3	IQR
1- Number of fatal accidents that occur due to pilotage problems per flight hour	1a, 2a, 3a, 4a	86	1,70	5,97	5,5	7	7	1,5
2- Number of fatal accidents that occur due to pilotage problems per takeoff	1a, 2a, 3a, 4a	90	1,62	6,07	6	7	7	1
3- Number of incidents per flight hour	1a, 2a, 3a, 4b	83	1,35	5,79	5	6	7	2
4- Number of incidents per takeoff	1a, 2a, 3a, 4b	83	1,45	5,79	5	6	7	2
5*- Number of fatalities due to pilotage problems per takeoff	1a, 2a, 3a, 4a	62	1,98	5,24	4	6	7	3
6- Number of fatal injuries due to pilotage problems per takeoff	1a, 2a, 3a, 4a	72	1,80	5,59	4	6	7	3
7- Number of serious injuries due to pilotage problems per takeoff	1a, 2a, 3a, 4a	72	1,68	5,62	4	6	7	3
8- Number of cases where meteorological conditions are not evaluated in solo flights per takeoff	1b, 2b, 3b, 4b	76	1,72	5,62	4,5	7	7	2,5
9*- Number of cases where documents (related to flight and flight training) appropriate for aircraft type are not used per year	1b, 2b, 3b, 4b	69	2,06	5,24	3	6	7	4
10- Number of bird strikes per landing and takeoff	0	86	1,48	5,76	5	6	7	2
11- Number of runway excursions per landing	1a, 2a, 3a, 4b	86	1,35	5,90	5	6	7	2
12- Number of air misses per takeoff	1a, 2a, 3a, 4b	86	1,41	5,93	5	6	7	2
13*- Number of radio communication (technical) failures per flight hour	0	69	1,86	5,03	3,5	5	7	3,5
14*- Number of using the wrong taxiway per takeoff	1a, 2a, 3a, 4b	66	1,98	5,14	4	6	7	3
15- Number of wing strikes during landing per landing	1a, 2a, 3a, 4b	79	1,49	5,93	5,5	7	7	1,5
16- Number of non-compliance with ATC instructions per takeoff	1b, 2b, 3b, 4b	86	1,45	6,03	5,5	7	7	1,5
17*- Number of aborted flights due to psychoactive substance use per takeoff	1b, 2b, 3b, 4b	72	2,15	5,14	3,5	6	7	3,5
18- Number of non-compliance with SOPs per takeoff	1b, 2b, 3b, 4b	83	1,46	5,93	5	7	7	2
19- Number of wing strikes during takeoff per takeoff	1a, 2a, 3a, 4b	76	1,75	5,72	4,5	7	7	2,5
20- Number of aircraft collisions with ground vehicles per takeoff	0	79	1,95	5,34	5	6	7	2
21- Number of critical fuel situations per takeoff	1a, 2a, 3a, 4b	83	1,40	6,03	5,5	7	7	1,5
22- Number of critical fuel situations per flight hour	1a, 2a, 3a, 4b	79	1,50	5,76	5	6	7	2
23*- Number of overshoot per landing	1b, 2b, 3b, 4b	72	1,69	5,31	4	6	7	3
24- Number of cases where safety measures are not taken during refueling per takeoff	0	79	1,70	5,66	5	6	7	2
25- Number of cases where the centerline is missed during takeoff per takeoff	1a, 2a, 3a, 4b	79	1,72	5,55	5	6	7	2
26*- Number of cases where the runway track is missed during landing per landing	1a, 2a, 3a, 4b	66	2,04	5,17	4	6	7	3
27- Number of non-compliance with limitations specified in the Pilot's Operating Handbook per flight hour	1b, 2b, 3b, 4b	86	1,45	6,03	5	7	7	2
28- Number of damage to aircraft structure during landing per landing	1a, 2a, 3a, 4b	79	1,37	5,90	5	6	7	2
29*- Number of aborted flights due to FODs during takeoff per takeoff	0	72	1,76	5,45	4	6	7	3
30*- Number of hits by aircraft propeller to staff per takeoff	1a, 2a, 3a, 4a	69	2,06	5,34	3,5	6	7	3,5
31*- Number of incidents of flight crew fainting during flight per flight hour	1a, 2a, 3a, 4b	66	2,06	5,24	3	6	7	4

32*- Number of cases where flight crew reports living FODs on the taxiway per takeoff	0	69	1,90	5,21	4	6	7	3
33- Number of tire blowout per landing	1a, 2a, 3a, 4b	90	1,28	6,17	6	7	7	1
34- Number of hard landings per landing	1a, 2a, 3a, 4b	93	1,11	6,21	6	7	7	1
35- Number of cases where improper navigational aids are used during flight per takeoff	1b, 2b, 3b, 4b	79	1,74	5,48	5	6	7	2
36- Number of flights without debriefing at the end per landing	1b, 2b, 3b, 4b	72	1,90	5,10	4	6	6	2
37- Number of cases where a pilot operates a night flight without taking night flight training per takeoff	1b, 2b, 3b, 4b	72	2,24	5,66	4	7	7	3
38- Number of landing gear failures due to technical problems per landing	0	83	1,90	5,79	5,5	7	7	1,5
39- Number of tail strikes on the runway during takeoff per takeoff	1a, 2a, 3a, 4b	83	1,55	5,97	5,5	7	7	1,5
40- Number of tail strikes on the runway during landing per landing	1a, 2a, 3a, 4b	93	1,01	6,34	6	7	7	1
41- Number of engine (technical) failures per flight hour	0	86	1,75	5,86	6	6	7	1
42- Number of cases where takeoff is aborted at high speeds per takeoff	1a, 2a, 3a, 4b	90	1,40	6,03	5,5	7	7	1,5
43- Number of cases where meteorological conditions in the airport of arrival are not examined per takeoff	1b, 2b, 3b, 4b	83	1,85	5,86	6	7	7	1
44- Number of cases where NOTAMs related to the flight are not examined per takeoff	1b, 2b, 3b, 4b	83	1,85	5,86	6	7	7	1
45*- Number of cases where ground vehicles enter the PAT area without any permission during flight training per takeoff	0	69	2,34	5,03	3	6	7	4
46- Number of cases where the engine is over speed per flight hour	0	83	1,72	5,66	5	6	7	2
47- Number of SEP flights operated when meteorological conditions do not meet VFR or VMC requirements	1b, 2b, 3b, 4b	86	1,64	5,97	5	7	7	2
48- Number of emergencies reported per takeoff		79	1,66	5,79	5	7	7	2
49- Number of cases regarding not checking around in ground movements per landing and takeoff	1b, 2b, 3b, 4b	79	1,82	5,41	5	6	7	2
50*- Number of cases where the taxi speed limit is exceeded per landing and takeoff	1b, 2b, 3b, 4b	69	1,88	5,21	4	6	7	3
51- Number of getting lost during the flight per flight hour	1a, 2a, 3a, 4b	79	1,59	5,90	5,5	7	7	1,5
52- Number of emergency landing on terrain due to technical failure per takeoff	0	79	1,44	6,07	6	7	7	1
53- Number of cases where VFR flight conditions are lost per flight hour	1a, 2a, 3a, 4b	79	1,50	5,76	5	6	7	2
54- Number of propeller strikes per landing and takeoff	1a, 2a, 3a, 4a	83	1,48	6,03	6	7	7	1
55- Number of incidents in a flight reported by the ATC tower per landing and takeoff	1a, 2a, 3a, 4b	79	1,72	5,66	5	6	7	2
56- Number of cases where the VFR minima (e.g. flying into clouds, getting too close to the ground) are exceeded during flight training (e.g. turns, stalls, slow flight) per flight hour	1b, 2b, 3b, 4b	83	1,58	6,00	6	7	7	1
57*- Number of flight cancellations due to insufficient knowledge per takeoff	1b, 2b, 3b, 4b	66	2,26	4,76	3	6	7	4
58- Number of mid-air collisions per landing and takeoff	1a, 2a, 3a, 4a	66	1,94	5,52	4	7	7	3
59- Number of mid-air collisions per flight hour	1a, 2a, 3a, 4a	76	1,63	5,79	4,5	7	7	2,5
60*- The number of not adequate training concerning the phases of flight training.	1b, 2b, 3b, 4b	52	2,10	4,52	2,5	5	6	3,5
61- Number of cases where pre-flight control is not performed per takeoff	1b, 2b, 3b, 4b	69	1,86	5,52	4	7	7	3
62*- Number of fire incidents per flight hour		72	2,08	5,45	4	6	7	3
63- Number of NOTAM violations per takeoff	1b, 2b, 3b, 4b	83	1,68	5,97	5	7	7	2
64- Number of cases when crosswind assessment is not	1b, 2b, 3b, 4b	79	1,94	5,28	5	6	7	2

performed per landing								
65*- Number of cases where aircraft hits a ground vehicle per landing and takeoff	1a, 2a, 3a, 4a	72	1,86	5,48	4	6	7	3
66*- Number of stalls per flight hour	1a, 2a, 3a, 4b	55	2,26	4,66	2,5	6	7	4,5
67*- Number of stalls per landing and takeoff	1a, 2a, 3a, 4b	55	2,28	4,72	3	5	7	4
68*- Number of flight control system (technical) failures per landing and takeoff	0	72	2,16	5,45	4	7	7	3
69- Number of flights exceeding MTO limits per takeoff	1b, 2b, 3b, 4b	76	1,73	5,69	4,5	6	7	2,5
70*- Number of missed approaches per landing	1a, 2a, 3a, 4b	45	2,01	3,97	2	4	5,5	3,5
71- Number of cases where reporting is made at the wrong altitude and/or point per landing and takeoff	1b, 2b, 3b, 4b	76	1,87	5,28	4,5	6	7	2,5
72*- Number of cases where living FODs are reported on the runway per landing and takeoff	0	72	1,82	5,38	4	6	7	3
73*- Number of cases where non-living FODs are reported on the runway per landing and takeoff	0	72	1,88	5,34	4	6	7	3
74*- Number of flaming incidents during refueling per takeoff	0	69	2,18	5,38	4	7	7	3
75- Number of incidents reported by crew per landing and takeoff	1a, 2a, 3a, 4b	72	1,66	5,52	4	6	7	3
76- Number of hazards related to a flight reported by crew per landing and takeoff	1b, 2b, 3b, 4b	83	1,48	5,86	5	6	7	2
77- Number of cases where ATC instructions are misunderstood per landing and takeoff	1a, 2a, 3a, 4b	86	1,65	5,90	6	6	7	1
78- Number of students incapacitated per landing and takeoff	1a, 2a, 3a, 4b	76	1,90	5,41	4,5	6	7	2,5
79- Number of flight instructor incapacitated per landing and takeoff	1a, 2a, 3a, 4b	76	1,96	5,48	4,5	6	7	2,5
80- Number of cases where the altitude limit is exceeded per landing and takeoff	1b, 2b, 3b, 4b	72	1,81	5,52	4	6	7	3
81- Number of system (technical) failures that compromise flight safety per landing and takeoff	0	83	1,70	5,79	5	7	7	2
82- Number of landings on the wrong runway per landing	1a, 2a, 3a, 4b	83	1,59	6,03	6	7	7	1
83- Number of takeoffs from the wrong runway per takeoff	1a, 2a, 3a, 4b	72	2,11	5,55	4	7	7	3
84*- Number of cases where the engine is started without authorization per takeoff	1b, 2b, 3b, 4b	66	2,15	5,00	4	6	7	3
85- Number of violations of control or flight training area per flight hour	1b, 2b, 3b, 4b	86	1,54	5,79	5	6	7	2
86- Number of unauthorized entries to the runway per takeoff	1a, 2a, 3a, 4b	86	1,54	6,10	6	7	7	1
87- Number of landing gear (technical) failures per landing	0	76	1,96	5,59	4,5	7	7	2,5
88- Number of non-compliance with approaching and/or landing instructions in VFR flights per landing	1b, 2b, 3b, 4b	76	1,80	5,79	4,5	7	7	2,5
89- Number of cases where the pilot forgets to lower the landing gear per landing	1a, 2a, 3a, 4b	79	1,92	5,79	5	7	7	2
90*- The number of quality audit finding which is not closed within the specified time per landing and takeoff.	1b, 2b, 3b, 4b	62	2,33	4,72	2,5	6	7	4,5
91- Number of self-reports of error by flight crew per landing and takeoff	1b, 2b, 3b, 4b	83	1,71	5,72	5	6	7	2
92*- Degree of strong positive safety culture per year	1b, 2b, 3b, 4b	69	1,78	5,45	4	6	7	3
93*- Degree of support provided by senior management for safety enhancement per year	1b, 2b, 3b, 4b	66	1,90	5,14	4	5	7	3
94*- Degree of safety commitment of flight personnel per year	1b, 2b, 3b, 4b	62	2,00	5,07	4	5	7	3
95*- Degree of safety commitment of senior management per year	1b, 2b, 3b, 4b	66	1,97	5,17	4	6	7	3
96- Number of cases where takeoff running begins without takeoff authorization per takeoff	1b, 2b, 3b, 4b	79	1,68	5,79	5	7	7	2

97*- Number of wrong flap configurations in takeoff per takeoff	1b, 2b, 3b, 4b	72	1,74	5,41	4	6	7	3
98*- Number of deviations from the flight route due to a navigation error per flight hour	1b, 2b, 3b, 4b	62	2,07	5,00	4	6	7	3

D. Findings of the Fourth Delphi Round

In the fourth Delphi round, 66 SPIs obtained at the end of the third round were resent to experts without any revisions, who were asked to reassess the items in consideration of responses in the preceding round and general responses of the panel. After a Delphi round that took about 3.5 months, the panel reached a consensus on 64 out of 66 SPIs. Only two SPIs were eliminated in this round. This shows that the panel members came to a strong consensus on the remaining 64 SPIs given that they reached an agreement on these 64 items twice at different times. Thus, it may be concluded that the fourth Delphi round produced a final list consisting of 64 SPIs. The SPIs eliminated in this round were the following:

- 7- Number of cases where meteorological conditions were not evaluated in solo flights per takeoff
- 25- Number of flights without debriefing at the end per landing

The two SPIs eliminated are leading, activity-oriented and proactive indicators that intend to check whether safety enhancement activities have been carried out. The fact that these indicators are eliminated is a factor that reduces the success of SPI design. The list consisting of 64 items was analyzed in view of perfection and validity criteria. The analysis results are presented in Table 10. Out of 64 SPIs, 21 (32.8%) are classified as perfect indicators. Eight SPIs, which were believed to be invalid since the beginning of research, were included in the list. This is a finding that compromises the success of SPI design.

Table 10. Analysis of the Final List of SPIs

Categorization of SPIs	SPIs in the Relevant Category		SPIs Not Included in the Relevant Category		Total
	Frequency	Rate (%)	Frequency	Rate (%)	
Perfect SPIs	21	32.81	43	67.19	64
SPIs not eliminated although they need to be**	8	12.50	56	87.50	64

Table 11 present the types of 56 final SPIs, and excludes eight SPIs that needed to be eliminated but were not. The fact that 85% of SPIs are low-consequence indicators points to high success of indicator design process. However, the rate of leading, activity-oriented and reactive SPIs remained at 37.5%. The rate is quite low and jeopardizes the success of SPI design.

Table 11. Types of Final SPIs

Indicator Group	Codes	Type	Number of Indicators	%
1 Leading Lagging	1a	Lagging	35	62.50
	1b	Leading	21	37.50
		Total	56	100
2 Outcome Activity	2a	Outcome	35	62.50
	2b	Activities	21	37.50
		Total	56	100
3 Reactive Proactive	3a	Reactive	35	62.50
	3b	Proactive	21	37.50
		Total	56	100
4 Low Level High Level	4a	High Level	8	14.28
	4b	Low Level	48	85.72
		Total	56	100

E. Findings of the Fifth Delphi Round

At the end of four Delphi rounds, the experts came to an agreement on 64 final SPIs. In the last round of the Delphi study, the experts were asked to rate 64 indicators from 0 to 100 with regard to their level of importance in measuring safety accurately. The total score was divided into the number of participants to obtain the mean values. Table 12 provides the list of SPIs ranked in the order of importance, mean and standard deviation values of participants' scores, and types and validity status of SPIs. The '0' value in the column of 'type of SPI' indicates that the given SPI is invalid.

Table 12. Ranking of Final SPIs with Regard to Importance

	Safety Performance Indicators (SPIs)	Mean	SD	Type of SPI
1	Number of non-compliance with limitations specified in the Pilot's Operating Handbook per flight hour	88,65	18,69	1b, 2b, 3b, 4b
2	Number of tire blowout per landing	86,35	22,34	1a, 2a, 3a, 4b
3	Number of hard landings per landing	83,15	20,37	1a, 2a, 3a, 4b
4	Number of damage to aircraft structure during landing per landing	82,50	24,90	1a, 2a, 3a, 4b
5	Number of runway excursions per landing	82,25	23,69	1a, 2a, 3a, 4b
6	Number of getting lost during the flight per flight hour	81,95	24,42	1a, 2a, 3a, 4b
7	Number of flight instructor incapacitated per landing and takeoff	81,85	29,07	1a, 2a, 3a, 4b
8	Number of serious injuries due to pilotage problems per takeoff	81,60	28,83	1a, 2a, 3a, 4a
9	Number of bird strikes per landing and takeoff	80,95	20,99	0
10	Number of engine (technical) failures per flight hour	80,40	27,38	0
11	Number of fatal accidents that occur due to pilotage problems per flight hour	79,95	21,25	1a, 2a, 3a, 4a
12	Number of air misses per takeoff	79,90	29,59	1a, 2a, 3a, 4b
13	Number of incidents per takeoff	79,80	23,71	1a, 2a, 3a, 4b
14	Number of cases where the pilot forgets to lower the landing gear per landing	79,40	29,34	1a, 2a, 3a, 4b
15	Number of wing strikes during landing per landing	79,35	24,22	1a, 2a, 3a, 4b
16	Number of cases where the centerline is missed during takeoff per takeoff	79,30	23,74	1a, 2a, 3a, 4b
17	Number of tail strikes on the runway during landing per landing	79,20	29,51	1a, 2a, 3a, 4b
18	Number of non-compliance with ATC instructions per takeoff	79,00	27,14	1b, 2b, 3b, 4b
19	Number of critical fuel situations per takeoff	78,95	26,45	1a, 2a, 3a, 4b
20	Number of non-compliance with SOPs per takeoff	78,40	30,43	1b, 2b, 3b, 4b
21	Number of incidents per flight hour	78,35	18,89	1a, 2a, 3a, 4b
22	Number of tail strikes on the runway during takeoff per takeoff	78,35	29,66	1a, 2a, 3a, 4b
23	Number of propeller strikes per landing and takeoff	77,95	32,07	1a, 2a, 3a, 4a
24	Number of fatal accidents that occur due to pilotage problems per takeoff	77,85	31,01	1a, 2a, 3a, 4a

25	Number of fatal injuries due to pilotage problems per takeoff	77,85	31,38	1a, 2a, 3a, 4a
26	Number of cases where pre-flight control is not performed per takeoff	77,65	33,62	1b, 2b, 3b, 4b
27	Number of cases where NOTAMs related to the flight are not examined per takeoff	77,55	29,57	1b, 2b, 3b, 4b
28	Number of cases where the engine is over speed per flight hour	77,55	26,88	1a, 2a, 3a, 4b
29	Number of cases where a pilot operates a night flight without taking night flight training per takeoff	77,05	32,90	1b, 2b, 3b, 4b
30	Number of NOTAM violations per takeoff	76,85	35,93	1b, 2b, 3b, 4b
31	Number of flights exceeding MTO limits per takeoff	76,85	34,95	1b, 2b, 3b, 4b
32	Number of SEP flights operated when meteorological conditions do not meet VFR or VMC requirements	76,70	28,93	1b, 2b, 3b, 4b
33	Number of unauthorized entries to the runway per takeoff	76,10	36,61	1a, 2a, 3a, 4b
34	Number of cases where meteorological conditions in the airport of arrival are not examined per takeoff	76,00	30,20	1b, 2b, 3b, 4b
35	Number of emergencies reported per takeoff	74,30	31,88	1a, 2a, 3a, 4a
36	Number of cases where improper navigational aids are used during flight per takeoff	74,15	30,59	1b, 2b, 3b, 4b
37	Number of wing strikes during takeoff per takeoff	73,55	30,22	1a, 2a, 3a, 4b
38	Number of cases where the VFR minima (e.g. flying into clouds, getting too close to the ground) are exceeded during flight training (e.g. turns, stalls, slow flight) per flight hour	73,40	30,34	1b, 2b, 3b, 4b
39	Number of landing gear failures due to technical problems per landing	73,35	31,23	0
40	Number of hazards related to a flight reported by crew per landing and takeoff	73,35	32,47	1b, 2b, 3b, 4b
41	Number of cases where safety measures are not taken during refueling per takeoff	73,30	32,67	0
42	Number of cases when crosswind assessment is not performed per landing	72,05	32,01	1b, 2b, 3b, 4b
43	Number of critical fuel situations per flight hour	72,00	22,40	1a, 2a, 3a, 4b
44	Number of cases where VFR flight conditions are lost per flight hour	72,00	29,17	1a, 2a, 3a, 4b
45	Number of incidents in a flight reported by the ATC tower per landing and takeoff	71,80	32,61	1a, 2a, 3a, 4b
46	Number of system (technical) failures that compromise flight safety per landing and takeoff	71,75	36,59	0
47	Number of students incapacitated per landing and takeoff	71,55	31,58	1a, 2a, 3a, 4b
48	Number of violations of control or flight training area per flight hour	71,40	30,71	1b, 2b, 3b, 4b
49	Number of landings on the wrong runway per landing	71,20	35,62	1a, 2a, 3a, 4b
50	Number of cases where takeoff running begins without takeoff authorization per takeoff	70,90	37,74	1b, 2b, 3b, 4b
51	Number of non-compliance with approaching and/or landing instructions in VFR flights per landing	70,65	34,36	1b, 2b, 3b, 4b
52	Number of emergency landing on terrain due to technical failure per takeoff	69,95	37,46	0
53	Number of mid-air collisions per landing and takeoff	69,75	36,70	1a, 2a, 3a, 4a
54	Number of incidents reported by crew per landing and takeoff	69,40	33,47	1a, 2a, 3a, 4b
55	Number of cases where the altitude limit is exceeded per landing and takeoff	69,20	34,89	1b, 2b, 3b, 4b
56	Number of takeoffs from the wrong runway per takeoff	69,10	36,09	1a, 2a, 3a, 4b
57	Number of mid-air collisions per flight hour	68,80	32,93	1a, 2a, 3a, 4a
58	Number of landing gear (technical) failures per landing	68,50	34,62	0
59	Number of self-reports of error by flight crew per landing and takeoff	68,20	35,70	1b, 2b, 3b, 4b
60	Number of cases where takeoff is aborted at high speeds per takeoff	66,25	33,19	1a, 2a, 3a, 4b
61	Number of cases regarding not checking around in ground movements per landing and takeoff	63,95	33,01	1b, 2b, 3b, 4b
62	Number of cases where ATC instructions are misunderstood per landing and takeoff	62,85	31,86	1a, 2a, 3a, 4b
63	Number of aircraft collisions with ground vehicles per takeoff	62,30	35,12	0
64	Number of cases where reporting is made at the wrong altitude and/or point per landing and takeoff	60,85	32,53	1b, 2b, 3b, 4b

"Number of incidents of non-compliance with limits specified in the Pilot's Operating Handbook per flight hour" received the highest score in the fifth Delphi round, raking the first in the list of importance. This is an SPI that falls under the category of perfect SPIs. This is a positive finding that points to the success of SPI design. However, among the first 16 SPIs in the list of importance, there is only one perfect SPI. On the other hand, there are two invalid SPIs among the first 10 indicators in the list of importance. Participants considered important and gave high scores to two SPIs that were basically invalid. "Number of bird strikes per landing and takeoff" is an SPI that measures safety performance of an airport, and "number of engine (technical) failures per flight hour" is an indicator that intends to measure safety performance of maintenance activities.

Table 13. Distribution of Perfect SPIs

Percentile Range	Number and Rank of SPIs	Number of SPIs in the Range	Total Number of Perfect SPIs
25	16	1	1
50	32	8	9
75	48	6	15
100	64	6	21

Table 13 shows the distribution of perfect SPIs. Out of 64 SPIs, there is one perfect SPI in the first quartile range, and there are eight SPIs in the second quartile range. These findings may indicate that participants see perfect SPIs neither unimportant nor very important.

CONCLUSION AND RECOMMENDATIONS

This study based on the Delphi method yielded a list of 64 SPIs on which an expert panel came to a consensus. The most important result of the study is that production parameters are not taken into consideration in SPI design. Risk is nevertheless affected by production amount. Measurement results are inaccurate if there is no specific information on the field of activity in which safety performance measurement is conducted, what is produced in the given field, and how to measure the production validly, and if unwanted incident and process management is not associated with aforementioned factors. Inaccurate safety performance measurement render SMS invalid, and results in erroneous decision making.

Another significant result of the study is that experts tend to design SPIs that measure safety performance in other fields of activity rather than dealing only with flight training. This is also likely to produce inaccurate measurements, rendering SPIs invalid.

Another result of this study is that experts generally tend to focus on lagging and outcome-oriented reactive indicators rather than on leading, activity-oriented and proactive indicators. This is mainly because experts have limited knowledge of the latter and their functions. SMS, the product of performance-based approach to safety, is a new system in Turkey and the world. That is why there is limited information about these indicators. The regulation-based approach prevailed in aviation safety management from the establishment of the ICAO and national civil aviation authorities to the end of 2000s. As reactive characteristics are dominant in this approach, it is acceptable that proactive indicators have not been popular so far. However, if the aviation sector wants SMS to achieve the expected benefits, flight training sector needs to be informed about leading and proactive indicators immediately.

REFERENCES

ATAG. (2012, Mart 22). *Towards Sustainable Aviation*. Geneva, Switzerland.

Cova, T. J., & Conger, S. (2004). Transportation Hazards. M. (ed.) içinde, *Handbook of Transportation Engineering* (s. 17.1-17.24). New York: McGraw Hill.

Daft, R. L. (2008). *Management. (8.basım)*, Southwestern: Thomson.

Dyreborg, J. (2009). The causal relation between lead and lag indicators. *Safety Science*, 47 (4), 474-475.

Emil, M. F., Yılmaz, H. H., & Rijkeghem, C. V. (2005). *Kamu Borcu Nasıl Oluştu? Bu Noktaya Nasıl Gelindi?* İstanbul: Tesev Yayınları Türkiye-Avrupa Birliği İlişkileri Serisi.

Erikson, S. G. (2009). Performance indicators. *Safety Science*, 47 (4), 468.

Gerede, E. (2015a). A study of challenges to the success of the safety management system in aircraft maintenance organizations in Turkey. *Safety Science*, 73, 106-116.

Gerede, E. (2015b). A qualitative study on the exploration of challenges to the implementation of the safety management system in aircraft maintenance organizations in Turkey. *Journal of Air Transport Management*, 47, 230-240.

Gençtürk, e., & Akbaş, y. (2013). Sosyal Bilgiler Öğretmenliği Coğrafya Alan Standartlarının Belirlenmesi: Delphi Tekniği Uygulaması. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *33*(2).

Gerede, E. (2016). Emniyet Performansının Ölçülmesi ve İzlenmesi. E. Gerede içinde, *Havacılık Emniyeti* (s. 158-184). Eskişehir: Anadolu Üniversitesi Yayınları.

Hale, A. (2009). Why Safety Performance Indicators? *Safety Science, 47*(4), 479-480.

Hatcher, T., & Colton, S. (2007). Using the internet to improve HRD research: The case of the web-based Delphi research technique to achieve content validity of an HRD-oriented measurement. *Journal of European Industrial Training, 31*(7), 570-587. doi:10.1108/03090590710820060

Heavisides, J. and Thompson, K. (2011). Investigating the Application of Leading and Lagging Indicators to the rail industry: Measuring Safety Performance. *Rail Safety and Standards Board.*

Hollnagel, E. and Woods, D. D. (2006). *Epilogue – Resilience Engineering Precepts, içinde: Resilience engineering: Concepts and Precepts*, Editor: E. Hollnagel, D. D. Woodsand, N. Leveson, Aldershot: Ashgate.

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi Technique: Making Sense of Consensus. *Practical Assesment, Research & Evaluation, 12*(10), 1-8.

Hung, H.-L., Altschuld, J. W., & Lee, Y.-F. (2008). Methodological and conceptual issues confronting a cross-country Delphi study of educational program evalutation. *Evaluation and Program Planning, 31*, 191-198.

IATA. (2007). *Aviation Economic Benefits*. (M. Smyth, & B. Pearce, Dü) Switzerland: IATA.

ICAO, (2009). *Safety management manual. Doc 9859 AN/474, (2.basım)*, Montreal, Quebec.

ICAO, (2013a). *Safety management manual. Doc 9859 AN/474 (3.basım)*, Montreal, Quebec.

ICAO, (2013b). *Safety management. Annex 19 to the Convention on International Civil Aviation*, Montreal, Quebec.

Köklü, N. (1995). Tutumların Ölçülmesi ve Likert Tipi Ölçeklerde Kullanılan Seçenekler. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 28*(2), 81-93.

Linstone, H. A., & Turoff, M. (2002). *The Delphi Method Techniques and Applications*. Ocak 18, 2017 tarihinde https://pdfs.semanticscholar.org/8634/72a67f5bdc67e4782306efd883fca23e3a3d.pdf adresinden alındı

Linz, M. (2012). Scenarios for the aviation industry: A Delphi-based analysis for 2025. *Journal of Air Transport Management*, *22*, 28-35.

Lohmann, G., Albers, S., Koch, B., & Pavlovich, K. (2009). From Hub to Tourist Destination-An explorative study of Singapore and Dubai's Aviation-Based Transformation. *Journal of Air Transportation*, 205-211.

Löfmark, A., & Mårtensson, G. (2017). Validation of the tool assessment of clinical education (AssCE): A study using Delphi method and clinical experts. *Nurse Education Today*, 82-86.

O'Connor, W. E. (2001). *An Introduction to Airline Economics Sixth Edition.* London: Praeger.

OECD, (2008). Guidance on developing safety performance indicators related to chemical accident prevention, preparedness and response: For Industry. (2.basım), *OECD Environment, Health and Safety Publications, Series on Chemical Accidents No. 19*. OECD Publications: Paris.

Øien, K., Utne, I.B., Tinmannsvik, R.K. and Massaiu, S. (2011). Building Safety Indicators: Part 2 –Application, practices and results. *Safety Science*, 49 (2), 162-171.

Powell, C. (2003). The Delphi Technique: myths and realities. *Journal of Advanced Nursing, 41*(4), 349-382.

Reiman, T. and Pietikäinen, E. (2012). Leading indicators of system safety – Monitoring and driving the organizational safety potential. *Safety Science*, 50 (10), 1993-2000.

Saçaklıoğlu, F., Karababa, A. O., Döner, B., Özakbaş, E. E., Kitapçıoğlu, G., Kandiloğlu, G., . . . Gürsoy, Ş. T. (2005). *Tıp Eğitiminde Müfredatın Geliştirilmesi İçin Toplumumuzun Öncelikli Sağlık Sorunlarının Belirlenmesi.* Ege Üniversitesi Tıp Fakültesi , Halk Sağlığı Anabilim Dalı. İzmir: Ege Üniversitesi Basımevi.

Safety Management International Collaboration Group, (2013). *Measuring Safety Performance: Guidelines for Service Providers.*

Sharkey, S. B., & Sharples, A. Y. (2001). An approach to consensus building using the Delphi technique: developing a learning resource in mental health. *Nurse education today*, *21*(5), 398-408.

Skumolski, G. J., Hartman, F. T., & Kahn, J. (2007). The Delphi Method for Graduate Research. *Journal of Information Technology Education, 6*(1), 1-21.

Step Change in Safety, (2003). *Leading Performance Indicators: Guidance for Effective Use.* Aberdeen.

Şahin, A. E. (2001). Eğitim Araştırmalarında Delphi Tekniği ve Kullanımı. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20*, 215-220.

Şahin, A. E. (2010). Professional status of elementary teaching in Turkey: a Delphi study. *Teachers and Teaching: theory and practice, 16*(4), 437-459.

Terzi, H., & Gazioğlu, C. (2014). Pozitif Emniyet Kültürü Temel Öğelerini Esas Alan Kazasay Olayları (Near Miss) Raporlama Sistemi. *Dokuz Eylül Üniversitesi Denizcilik Fakültesi Dergisi, 6*(1), 23-58.

UK Health and Safety Executive, (2006). *Developing process safety indicators: a step-by-step guide for chemical and major hazard industries.*

Verstraeten, J. G., Roelen, A. L. C. and Speijker, L. J. P. (2014). *Safety performance indicators for system of organizations in aviation*. Aviation Safety and Certification of new Operations and Systems Grant Agreement No: 314299.