Available online at http://docs.lib.purdue.edu/jate

JATE

Journal of Aviation Technology and Engineering 9:2 (2020) 19–34

An HFACS Analysis of German F-104 Starfighter Accidents

Steven Esser and Hans-Joachim K. Ruff-Stahl

Embry-Riddle Aeronautical University

Abstract

From 1961 onwards, Germany acquired 916 Lockheed F-104 Starfighters, of which 292 aircraft crashed and 116 pilots lost their lives. The purpose of this research project was to find out why these aircraft crashed and whether the Starfighters crashed for reasons different from those for other military aircraft in Germany. Seventy-one German F-104 accidents between 1978 and 1986 were analyzed by reviewing the original accident files. A Human Factors Analysis and Classification System (HFACS) Level-1 analysis was used as methodology. It was found that more than 50% of the reviewed German F-104 accidents occurred due to technology and/or physical environment. More than half of the sample's accidents were engine related. It was concluded that the F-104 was indeed more accident-prone than other co-era types. Moreover, the J-79 engine was found to be a weak link in the F-104's safety record, and the Starfighter's unforgiving handling characteristics induced an elevated level of skill-based errors.

Keywords: F-104, Starfighter, HFACS, human factors

Introduction

In the late 1950s the German Ministry of Defense (MOD) was looking for a new combat aircraft to replace its aging fleet of Air Force and Navy fighters. Lockheed's F-104 Starfighter won the international competition in October 1958 after a prolonged selection process (Siano, 2016). This futuristic fighter was able to cruise at Mach 2 speeds, was packed full with latest electronics, and promised an unparalleled combat performance. Hence, the German MOD had high expectations for this brand new jet fighter.

However, these expectations quickly turned into a nightmare. Following its introduction into service in May 1960 an accident series developed, which rocked Germany's military, Germany's parliament, and Germany's public (Neu, 2004; Siano, 2016). The accident series peaked between February 1965 and July 1966, when on average once every two weeks an F-104 was lost. In this time frame 46 F-104 fighters crashed, killing 29 pilots. Even today this period is still infamously called Germany's "Starfighter Krise" [Starfighter crisis].

Although the crisis was eventually overcome and *accident rates* settled at or even below international rates, the *accident numbers* still grew at a stunning pace. By December 1968 the first 100 F-104s were lost; the 200th aircraft crashed in September 1975 and the last German F-104 crashed as late as 1989, while already out of operational service (Fischbach,

http://dx.doi.org/10.7771/2159-6670.1218

1998; Kropf, 2002). Despite being retired, sold, or scrapped more than 25 years ago, this 1950s jet aircraft to this date is still overshadowed by myths: in contrast to common press and public referrals as "widowmaker" or "flying coffin" and being perceived as a notoriously pilot-killing aircraft, its former pilots still hold it in highest esteem (Neu, 2004; Reis, 2012; Vogler, 1986). Nevertheless, of a total of 916 aircraft acquired by the German military 292 crashed or were damaged beyond repair during the F104's time in service from 1960 to 1991. This came at the cost of 116 dead pilots, including eight U.S. Air Force (USAF) exchange pilots (Fischbach, 1998; Kropf, 2002). Hence, the purpose of this project is to find out whether the former myth should be rejected, which lopsidedly blames the aircraft's technology, and which is challenged by contemporary findings on human error.

For almost three decades Germany's military, parliament, press, and public have frequently and repeatedly been affected by the Starfighter—obviously to varying degrees, but definitely unlike by any other piece of military equipment before or after (Neu, 2004). At the time of writing, searching the term "F-104 widowmaker" at Google leads to more than 150,000 results, which indicates a surprisingly strong echo from an era that ended almost three decades ago. Yet, surprisingly little scientific literature has been published on the Starfighter and a considerable gap in research exists on the Starfighter's accident causations. Although the F-104's flight safety issues were repeatedly addressed by the German parliament's defense committee, no scientific analysis of its accident causations exists, whatsoever. Consequently, no contemporary method of accident analysis and classification has ever been applied to the F-104 accidents either.

Broadly speaking, two very simplistic, contradicting, and extreme views on the F-104 prevail in publications (Reis, 2012). On the one hand, the myth of a poorly designed pilot-killing monster, as the press tended to report (Ein schöner Tod, 1982; Kauf von Schrott, 1969; Neu, 2004). And on the other hand, the sentimental pilot myth of a high-performing, flawless, but unforgiving beauty (Loy, 2011; Stiller, 1981; Vogler, 1986).

Research indicates that between 70% and 80% of aviation accidents can be attributed to human error (Dekker, 2002; Helmreich & Foushee, 2010; Shappell & Wiegmann, 2001). Moreover, research on accident causation also questions the simplistic explanations common in past decades (e.g., "pilot error") and suggests comprehensive system-analysis approaches instead (Reason, 2000; Shappell & Wiegmann, 2000). Hence, both prevailing views on the F-104 are in conflict with latest research on accident causations.

Consequently, the specific research question for this project was whether the German F-104s crashed for different reasons from those for *other* military aircraft. To answer this question, the last 71 German F-104 accidents, which occurred between 1978 and 1986, will first be analyzed, using the Human Factors Analysis and Classification System (HFACS) as a method. An organizational and managerial analysis of Germany's Air Force and Navy is beyond the scope of this paper; hence, the HFACS analysis is limited to its first level (unsafe acts).

The results of this analysis are expected to shed light on the exact causes of these German F-104 accidents and particularly on the relationship between human factor-caused accidents and technology-caused accidents.

Next, the HFACS results will be quantitatively compared with HFACS results of 72 USAF accidents. Due to a lack of sources on earlier decades' HFACS-coded USAF accidents, the time span from 1991 to 1997 had to be used as reference. This comparison will show whether the German F-104 accident causes differ significantly from accidents in U.S. military aviation. Additionally, the German F-104 accident rates and engine-caused accident rates will be compared with co-era USAF aircraft, such as the F-105 or F-106.

Only if the aircraft's advanced technology, labeled unmanageable by press and public, has caused significantly more accidents than what can be considered the norm would the "widowmaker" and "pilot-killer" myth be justified. For this reason this research project attempts to find out whether the F-104 crashed for technological and human factor causes different from those for aircraft in the U.S. Air Force.

This analysis is restricted to the last nine years of *operational* service in Germany's Air Force and Navy (1978 to 1986). The F-104 attained its bad reputation in the "Starfighter crisis" of 1965–1966, which would thus be an obvious choice for an F-104 HFACS analysis. But for meaningful results, a suitable control group would have to be an HFACS analysis of a similar aircraft (i.e., single-seat, single-engine supersonic fighter) from the same era, operated in the same environment (i.e., European weather, low-level flying)—and such an analysis simply does not exist; nor are the accident files of possible control groups (e.g. Danish F-100, Dutch or Canadian F-104 accidents in Europe) publicly accessible, which prevents the creation of an own data set. Hence, the available control group in terms of time span and volume (HFACS analysis of USAF accidents from 1991 to 1997) defined which German F-104 accidents had to be analyzed to achieve meaningful results (i.e., 1978 to 1986).

As mentioned before, an organizational and managerial analysis of Germany's Air Force and Navy is beyond the scope of this paper. Hence, the HFACS analysis is limited to its first level (unsafe acts). However, the second level (Preconditions for Unsafe Acts—Environmental Factors—Technological Environment and Physical Environment) will still be addressed in the results and discussion sections of this paper.

Shappell and Wiegmann (2004) have published an HFACS analysis of U.S. military aviation accidents, in which 72 USAF accidents in the time span of 1991 to 1997 have been analyzed. The preconditions for a perfect control group have been described above, but such a control group is simply not available. The available USAF control group has a similar volume and time span, but includes different types of aircraft operated in a different environment.

The exact percentage of the control group's data (see Tables 3 and 6) can only be estimated to within 2% of its true value. Unfortunately, contact with the authors of the mentioned paper could not be established. Hence, clarification could not be achieved.

The test group's HFACS coding was performed by a single person only. The coding and its reliability will be described in the Methodology section (subsections "HFACS Coding" and "Reliability"). Nevertheless, despite these limitations the results of this research are expected to be meaningful.

Literature Review

By 1955, only ten years after Germany's thorough defeat in World War II, the young Federal Republic of Germany was permitted to rebuild its armed forces. Germany had to do its part in a crucial effort for the common defense of Western Europe against the threat posed by the Soviet Union and the Warsaw Pact (Rebhan, 2006; Wettig, 1995).

However, Germany's military had to be rebuilt from scratch. The initial equipment procured by its military was mostly legacy and excess equipment from the United States, Canada, Great Britain, and even France. The initial outfit for the Air Force consisted of various types of trainer and cargo aircraft, while the combat units were equipped with F-84 and F-86 jets in different variants (Lemke, 2006; Rebhan, 2006). As the Korean War had shown, both types were already at the brink of obsolescence and the search for a successor started as early as 1956 (Lemke, 2006; Siano, 2016).

For the acquisition of a new fighter several factors had to be taken into consideration: (a) military requirements; (b) political requirements, and (c) industry requirements (Lemke, 2006; Schlieper, 1995; Siano, 2016). Initially, eleven fighter types in different stages of their development were being considered. Finally, the selection was narrowed down to three types: Grumman's F-11 Super Tiger, Dassault's Mirage III A, and Lockheed's F-104 (Siano, 2016).

The expected air threat in a future conflict with the Warsaw Pact was high-flying nuclear bombers like the TU-16 Badger, M-4 Bison, or TU-95 Bear, escorted by supersonic fighters such as the MiG-19 Farmer or the MiG-21 Fishbed. Simultaneously, smaller bombers like the Yak-25 or IL-28 might be attacking at lower altitudes. Not only were suitable surface-to-air missiles not yet available, but to make things worse Germany only had an east–west extension of 300 to 400 km. All this meant extremely short reaction times for North Atlantic Treaty Organization (NATO) defenders in the case of a surprise assault by Warsaw Pact forces. Hence, there was an evident need for a fighter with an unsurpassed climb rate, and a need for extremely high cruise speeds (Lemke, 2006; Schlieper, 1995).

Additionally, NATO's defense strategy at the time was MC 14/2 which aimed at maintaining peace by deterring Warsaw Pact aggression with the threat of massive nuclear retaliation (NATO, 1957). Hence, the German MOD considered its means as part of a credible deterrence would be a fighter capable of delivering nuclear ordnance (Siano, 2016).

Lastly, and mostly for financial reasons the German MOD was looking for a single multi-role fighter type instead of several different platforms. This multi-role fighter was supposed to equip the German Air Force as well as the Navy. It was supposed to be a pure air-to-air fighter, a reconnaissance platform, as well as a nuclear and conventional bomber—all in one airframe and without any reconfiguration to fulfill each role (Siano, 2016).

Two diametrically opposed political interests had to be recognized in the selection process. The first political interest involved France: after two world wars Germany now intended to tie the bonds to neighboring France by acquiring the new Mirage III jet. However, France refused to sell nuclear weapons to Germany. The second political interest involved the United States, which was the largest power in the Western Alliance—the United States had agreed reluctantly to the integration of nuclear weapons into a future German fighter. But the deal would only be closed under the condition that a U.S. product was bought, and under the condition of U.S. control over these weapons (Lemke, 2006; Siano, 2016).

At the end of World War II Germany possessed one of the largest and strongest aviation industries worldwide. But after the lost war a forced break of ten years was induced in which the industry was dismantled and specialists had to change their career paths to survive. In the meantime anything related to aviation and air war (e.g., airframes, engines, avionics, weapons, etc.) had seen unprecedented leaps forward. When the German aviation industry finally regained permission to design and build aircraft in 1955, it lacked knowledge and excellence in all sectors.

Both the German aviation industry and the German MOD intended to rebuild a modern and competitive German aviation industry. Therefore, both were looking for a comprehensive knowledge and technology transfer of a cutting-edge fighter design. This technology transfer should also include the establishment of a domestic fighter production line. Such a comprehensive and permissive contract was only offered by the Lockheed Aircraft Corporation for its F-104 fighter (Lemke, 2006; Siano, 2016).

In October 1958 the German MOD publicly announced that it would procure Lockheed's F-104G Starfighter. The G-version ("G" for Germany) was a further development of the existing F-104, specially tailored to meet Germany's requirements (Bowman, 2000; Kropf, 2002; Siano, 2016).

Lockheed's F-104 fighter made its maiden flight in April 1954. Unlike most fighter designs, this particular one did not originate from a military requirement but was a private venture of the Lockheed Aircraft Corporation. Lockheed's design chief had visited USAF fighter pilots during the Korean War, the latter being under the impression of encounters with the Soviet-built MiG-15. Asked for their demands for a new fighter design the most common replies were "superior speed" and "superior climb-rate." Upon the design chief's return to the United States, Lockheed's engineering team drafted a fighter to meet those demands. The result was the F-104 (Bowman, 2000; Kropf, 2002). The new design set multiple world records for speed and altitude, but to Lockheed's disappointment, the USAF only showed limited interest in the type (F-104, 1958).

The F-104's performance data and design features are impressive even by today's standards—let alone for a design of the 1950s. It was able to reach more than 55,000 feet and could cruise at Mach 2 speeds. Kropf (2002) mentions an unofficial world record, when in 1966 a twin-seat F-104F accelerated from a standstill on the runway to Mach 2 within 3.5 minutes. The design also included sophisticated avionic features: a stability augmentation system improved the jet's handling characteristics; a boundary layer control system utilized compressor bleed air to energize the airflow over the landing flaps, thus enabling reduced approach speeds; and an automatic pitch control system was incorporated, which included a stick shaker and a stick kicker to prevent pilots from exceeding angle of attack limits (Bowman, 2000; Kropf, 2002; USAF, 1960)

Germany demanded still more avionic systems to be built into its G-model. Among others, these included: an anti-skid system for the wheel brakes; an inertial navigation system to make the jet independent of external navigation aids; an auto pilot system was included to relieve the pilot; and a new multi-function radar was built into its nose. Lastly, two different bombing computers (M2 and dual timer) were incorporated for the respective delivery of conventional and nuclear ordnance (Bundesministerium der Verteidiung [BMVg], 1985; Lemke, 2006; Lockheed, 1960).

However, this cutting-edge performance came with a price tag. For instance, its distinct T-tail configuration offered increased performance over other tail designs, such as reduced drag and thus increased performance. It also provided improved stability and pitch control over a wide speed range. But this design came with a disadvantage, too. Most aircraft tend to lower their nose when exceeding the critical angle of attack (AOA) which usually alleviates the situation. The F-104's T-tail design in contrast induced the exact opposite behavior. When approaching the critical AOA, vortices from the aircraft's forward section would strike the T-tail, thus forcing a snap increase in AOA. This phenomenon, called "pitch-up," required a significant amount of altitude below the aircraft for successful recovery—if it was recoverable at all (Reaves, 1961). To enable safe aircraft operation and to prevent pitch-up from occurring in the first place, the mentioned automatic pitch control system was built into to the F-104. This system was not only AOA-dependent, but also measured the aircraft's pitch-change rate. Thus, a stick shaker would warn a pilot from impending critical situations, and if the situation aggravated further the "kicker" fired and manually pushed the control stick forward.

Consequently, most pilots recall that the F-104 was not a very forgiving aircraft. Rall (2004, p. 284) recalls a pondering test pilot saying, "it's an honest airplane. If you make a mistake, it will kill you." But not all judgments were this harsh. Vogler (1986, p. 10) recalls: "She was never moody, unpredictable, spiteful or even dangerous… She represents simply the peak of what a carefully selected and trained individual, in full possession of his mental and physical faculties, can master."

On the other hand, almost a third of all acquired German Starfighters crashed. This stunning attrition rate must be seen in relation to its flying hours (i.e., accident rate), in relation to other F-104 users, and in relation to co-era platforms.

It has been pointed out that by the end of its service life the German F-104 accident rates had settled to international standards in the vicinity of two accidents per 10,000 flight hours.

Table 1 shows the accident rates of co-era USAF fighters. It should be noted that the USAF F-104 accident rates rank by far on top.

Table 2 shows the engine-related accident rates of single-engine and twin-engine co-era USAF fighters. It should be noted that the USAF F-104 engine-related accidents rank by far on top, and that the F-4 with two J-79 engines ranks considerably lower.

Table 1
Co-era USAF fighter accident rates.

Aircraft type	Accident rate per 100,000 flight hours	Years in service
F-104	30.63	28
F-100	21.22	38
F-105	17.83	27
F-101	14.65	28
F-102	13.69	29
F-106	9.47	40
F-5	8.82	27

Note. Co-era USAF fighter accident rates per 100,000 flight hours. Even if modern types are reviewed the F-104 ranks by far on top in terms of accident rates in USAF service. Adapted from Lyons and Nace (2007).

Table 2
USAF engine-related accident rates (single-engine and twin-engine aircraft).

	Accident rate per 100,000 flight hours	Engine
Single-engine aircraft		
F-104	9.48	J-79
F-100	5.61	J-57
F-105	4.56	J-75
F-102	3.41	J-57
F-106	2.04	J-75
Twin-engine aircraft		
F-4	0.16	J-79
F-111	0.49	T F-30

Note. Co-era USAF engine-related accident rates per 100,000 flight hours. The F-104 ranks by far on top in terms of engine-related accident rates in USAF service. Adapted from USAF (2015a, 2015b).

All in all the existing literature indicates that the F-104 was on the upper end of what 1950s technology could achieve—and on the upper end of what human beings can master. Moreover, by 1960 aviation technology was fairly reliable. When the F-104G went into production it was already more likely that accidents were caused by the "human factor" than by technological flaws (Dekker, 2002; Helmreich & Foushee, 2010; Shappell & Wiegmann, 2001). Hence, a comprehensive accident analysis should encompass the F-104's technology and the environment in which it was operated, but also the human factor and the conditions under which the pilots operated this aircraft. The HFACS offers such a comprehensive taxonomy. Broadly speaking, accidents and human error can be viewed in two distinctly different ways: (a) in a person approach to human error or (b) in a system approach to human error.

The person approach to human error is a longstanding and die-hard view on unsafe acts. It usually focuses on the individual committing an unsafe act, i.e., the operator. This approach frequently identifies incompetence, inattention, or poor motivation as root causes for unsafe acts. Hence, this traditional view tends to see errors as individual shortcomings or even character flaws. Thereby the person approach to human error usually fosters a culture of naming and blaming. One evident effect of blame cultures is that they preclude safety cultures. In safety cultures individuals are encouraged to admit their errors to prevent others from repeating them.

In contrast, the system approach is a comprehensive approach to accidents and human errors, which tries to establish the whole picture. While it recognizes that human beings are fallible and will make mistakes, it also tries to identify the preconditions of unsafe acts, i.e., the conditions under which individuals operate (Dekker, 2002; Reason, 2000).

The aftermath of accidents or incidents usually provides ample information on an organization's safety culture and its approach to human error: if an organization's remedy aims at an individual alone (i.e., naming, blaming, shaming, retraining, disciplinary measures) it is a good indicator for the person approach to human errors. In contrast, an organization that analyses which safety feature has failed and whether organizational processes should be reviewed is likely to have adopted the system approach to accidents and human errors (Dekker, 2002; Reason, 2000).

Reason (1990) has advertised the latter view, i.e., the system approach to accidents and human error. He has suggested that in any organization human error can occur on four different levels. Each level represents a different safety layer with specific mechanisms to prevent failures. While historically accident investigators have often paid a disproportionate amount of attention to the lowest safety layer (i.e., Level-1 "Unsafe Acts"), Reason recognizes three more safety layers: Level-2 "Preconditions for Unsafe Acts"; Level-3 "Unsafe Supervision"; and Level-4 "Organizational Influences." This model is commonly referred to as the Swiss cheese model.

Reason recognizes that active failures occur exclusively on the lowest level (i.e., Level-1 "Unsafe Acts"), which is why this level usually receives the mentioned increased attention in the aftermath of events. The other three layers only contain latent failures, which may have existed unrecognized for years before a mishap. A "hole in a cheese slice" indicates a failure in the respective safety layer. If these holes are aligned in a way that an "event trajectory" is permitted, a mishap will be the result (Reason, 1990, 2000).

While the Swiss cheese model enables a look beyond the simplistic person approach to accidents and human error, it still is only useful for analyzing mishaps with hindsight. It is hardly useable for predicting trajectories and mishaps. Moreover, Reason did not specify what the "holes" in each slice are. Hence, this model lacked usability.

Shappell and Wiegmann (2000) developed a model called HFACS, which is a direct descendent of Reason's Swiss cheese model. This model recognizes the same four levels of failure, but specifies each level distinctly. Each level can be subdivided into different categories. For instance:

Level-1 "Unsafe Acts" can be broken into two different unsafe act *Categories,* which are "Errors" and "Violations." Errors and Violations in turn can be broken down into *Error types* (i.e., Decision Errors, Skill-Based Errors, and Perceptual Errors) and *Violation types* (i.e., Routine Violations, Exceptional Violations). Depending on required granularity, each error type can be further subdivided into different categories (Shappell & Wiegmann, 2000, 2001, 2003).

This model is a comprehensive system approach for the analysis of accidents. Shappell and Wiegmann (2003) have provided detailed guidance on the coding of Unsafe Acts (Level-1) and their preconditions (Level-2 to Level-4).

Hence, it is a very user-friendly model and enables coders to pinpoint where "holes in a cheese slice" have opened up.

However, as pointed out earlier, an organizational and managerial analysis of Germany's Air Force and Navy is beyond the scope of this paper. Hence, for the purpose of this research the HFACS analysis is limited to its first level (unsafe acts). Still, the second level (Preconditions for Unsafe Acts—Environmental Factors—Technological Environment and Physical Environment) will also be addressed in the results and discussion sections of this paper.

Errors can be defined as "mental or physical activities of individuals that fail to achieve their intended outcome." Violations in contrast can be defined as "the willful disregard for rules and regulations" (Shappell & Wiegmann, 2000, p. 3). What follows is an explanation of the different error types and violation types. While this is only an excerpt, Shappell and Wiegmann (2003) have listed numerous examples in checklist form to aid a correct coding.

Decision Errors represent behavior which proceeds according to a plan—yet, the plan was inappropriate for the given situation. Decision Errors are sometimes referred to as "honest mistakes," and can be further divided into three categories: (a) Procedural Errors, (b) Choice Errors, and (c) Problem Solving Errors.

Aviation examples for Procedural Errors are flawed "if–then" decisions, such as the application of an incorrect compressor stall clearing procedure, or flying an inappropriate maneuver.

An example for a Choice Error is the futile attempt to out-climb a thunderstorm, and ending up in the middle of it. Another example is a continued takeoff run instead of an abort, despite an afterburner failure.

Problem solving errors can occur, when a problem is not well understood, for instance due to a lack of information, training, or experience. For example, a misdiagnosed emergency situation would constitute a problem solving error (Shappell & Wiegmann, 2000, 2001, 2003).

Skill-Based Errors represent errors in which a task's demands exceed an individual's skills. Applied to aviation, Skill-Based Errors are failures of pilot basic flying skills, also known as "stick and rudder skills." An example of such a basic failure would be to stall an aircraft on approach. This error type can further be subdivided into three categories: (a) Attention Failures, (b) Memory Failures, and (c) Manner or Technique.

Two aviation examples for Attention Failures are (a) the breakdown of a pilot's visual scan, also known as cross-check, or (b) task fixation and channelized attention. Other examples are (c) distraction and (d) the inadvertent movement of flight controls. All of these Attention Failures could induce the mentioned stall on approach.

Examples for Memory Failures are omitted checklist steps or omitted steps of an emergency procedure.

Poor Operation Manners or Techniques refer to a pilot's inappropriate operation of the aircraft, such as manhandling it close to the ground, thus leading to a loss of control (Shappell & Wiegmann, 2000, 2001, 2003).

During flight pilots are exposed to different types of unnatural perceptions and sensations. These can lead to judgement errors, mostly when in impaired sensory conditions such as night flying or while flying in clouds. A well-known phenomenon in these conditions is spatial disorientation. However, spatial disorientation is no Perceptual Error. But a wrong pilot response to spatial disorientation constitutes a Perceptual Error. Misjudging the distance to the ground during a pull-out from a dive is another example for a Perceptual Error. Hence, Perceptual errors usually occur either in impaired sensory conditions or in high-task-load situations (Shappell & Wiegmann, 2000, 2001, 2003).

A Routine Violation is the deliberate and repeated violation of rules, such as frequently driving 20 km/h faster than permitted. An aviation example could be repeatedly and deliberately underflying minimum altitudes; another one would be the habit of underflying highway bridges. Routine Violations do not occur as a single event and are more often than not known to and tolerated by superiors, who fail to enforce the rules (Shappell & Wiegmann, 2000, 2001, 2003).

Exceptional Violations in contrast to Routine Violations are isolated departures of individuals who normally adhere to regulations and rules. An example could be to fly an aileron roll on a departure due to high spirits. An Exceptional Violation is neither characteristic for this individual, nor does it occur expectedly—which is why it is almost impossible to predict (Shappell & Wiegmann, 2000, 2001, 2003).

Hypotheses

As pointed out in the introduction, two simplistic and contradicting views on the F-104 prevail: on the one hand, the myth of the poorly designed pilot-killer; on the other hand, a somewhat sentimental pilot myth of the F-104 (Reis, 2012). Yet, both views on the F-104 are in conflict with current research. It is thus generally hypothesized that the F-104 is no outlier and that during its last years of operation the F-104 had similar accident trends to those of other military aircraft. Specifically, it is expected that a large number of those accidents can be attributed to human error and to the environment in which the F-104 was operated. Hence, the following specific hypotheses will be tested:

$H1_A$: The relative distribution of Errors and Violations contributing to German F-104 losses differs significantly from USAF accidents.

$H2_A$: The number of Decision Errors contributing to German F-104 losses differs significantly from USAF accidents.

$H3_A$: The number of Skill-Based Errors contributing to German F-104 losses differs significantly from USAF accidents.

$H4_A$: The number of Perceptual Errors contributing to German F-104 losses differs significantly from USAF accidents.

$H5_A$: The number of Violations (Routine and Exceptional) contributing to German F-104 losses differs significantly from USAF accidents.

$H6_A$: A strong association exists between error type and pilot flying hours on F-104.

Methodology

Empirical research was conducted to test these hypotheses in a quasi-experimental research design. The test group consists of a sample of 71 non-randomized German military Starfighter losses in the time frame from 1978 to 1986, which were analyzed with HFACS as method. These results were quantitatively compared with a control group of 72 non-randomized USAF accidents in a seven-year time frame from 1991 to 1997.

All German military aviation accident files are stored in the central archive of General Flugsicherheit der Bundeswehr (GenFlSichhBw) [Directorate of Aviation Safety, Federal Armed Forces]. These data are not accessible to the public. This research's authors were granted access to the original accident files by the Commanding Officer of GenFlSichhBw (Annex 2).

The USAF data have been analyzed and published by Shappell and Wiegmann (2004), by Lyons and Nace (2007) and by the USAF itself (2015a, 2015b).

GenFlSichhBw has granted this research under the following conditions (Annex 2): (a) Research will exclusively be performed at the Directorate of Aviation Safety; (b) the analysis will be anonymous; (c) copies may be made at the Directorate of Aviation Safety, but will not leave the building and will be kept on file; (d) Scientific publications must remain anonymous; and (e) the research's results will be made available to Director of Aviation Safety prior to publication.

Although 71 F-104 crashes were reviewed and coded, 37 had to be excluded from the hypothesis tests. The reason for this is two-fold: either no unsafe act occurred in that accident or the accident cause could not be identified. In 17 accidents the engine failed for technical reasons (technological environment). In 11 cases the engine failed after a bird strike (physical environment). Seven accidents had to be excluded for reasons other than the technological environment, such as controllability problems, stuck fuel, or gear malfunctions. Finally, one accident cause could not be identified and another aircraft vanished with its pilot into the Mediterranean Sea, leaving no trace.

It should be noted that the F-104 was neither equipped with a flight data recorder nor a cockpit voice recorder. Hence, those accident investigations were always challenging.

Although the hypothesis tests could only be conducted with a reduced test group of 34 cases, the research question can still be sufficiently addressed as to whether the German F-104s crashed for different reasons from those for *other* military aircraft. The USAF control group consists of 72 aircraft losses.

The coding into HFACS categories was performed according to the comprehensive and detailed descriptions of Shappell and Wiegmann (2000, 2001, 2003). The provided checklists were used to prevent omissions. Only the findings from the original investigation reports were taken into account and coded accordingly. No additions were made, nor were any findings left out.

In a very few special cases a translation of findings into contemporary concepts was required, though. For instance the concept of crew resource management (CRM) was not recognized in the original accident files, but is a separate Level-2 Personnel Factor *type*. For this translation the NOTECHS schema of CRM was applied (i.e., Leadership, Cooperation, Situation Awareness, Decision Making, and Communication). As an example, if inappropriate communication was one finding, it was thus coded as "CRM" finding (Flin et al., 2003). However, neither of those special cases was required for the HFACS Level-1 analysis.

Next, the HFACS Level-1 data were quantified to be tested against H1 to H6. These data are depicted together with the control group's data in Tables 3 and 6.

This research's author was the only HFACS coder of the accidents. The author is an active duty German Air Force officer, instructor pilot and weapons instructor on the Typhoon jet. As such he is an expert in aviation matters.

Moreover, HFACS has repeatedly been tested on the coding process reliability, its intra- as well as inter-rater reliability, and its consistency by Shappell & Wiegmann (2001, 2003), Hooper and O'Hare (2013), Ergai (2013), and Ergai et al. (2016) with excellent results, indicating the model's high reliability and consistency. In practical terms, therefore, even with only a single coder the data reliability can be regarded as sufficiently high.

The HFACS analysis provides both qualitative and quantitative data. To test hypotheses H1 and H6, nonparametric tests have been selected. H1 was tested with a chi-square goodness-of-fit test, while H6 was tested with a chi-square independence test (Weiss, 2016).

Since the research question for this project is whether the German F-104s crashed for different reasons from those for *other* military aircraft, the chi-square goodness-of-fit test is suitable to test the relative distribution of Errors and Violations between both samples.

The assumptions for a chi-square goodness-of-fit test are: (a) all expected frequencies are 1 or greater; (b) at most 20% of the expected frequencies are less than 5; and (c) simple random sample (Weiss, 2016).

The assumptions (a) and (b) have been fulfilled (Table 4). Assumption (c) has not been fulfilled, due to the quasi-experimental research design. The nine-year time frame and the intended sample size of 71 accidents were meant to induce randomization. Still, the method will be performed and its statistical power will be addressed.

Following the HFACS coding, Table 3 was constructed, which depicts the relative frequency of errors and violations for both test and control groups. Since accidents are rarely caused by a single factor only, the sum of errors and violations consequently exceeds 100%. However, a chi-square goodness-of-fit hypothesis test requires the sums of observed and expected relative frequency to add up to 100% each (Faul et al., 2007; Mayr et al., 2007; Weiss, 2016). Hence, each causation's relative frequency had to be factorized (0.6897 for USAF, 0.7556 for F-104) to achieve a sum of 100% (Table 4). It must be noted that the numerical *relation* of error and violation types remains unchanged by the factorization.

The chi-square goodness-of-fit calculations were performed with the StatCrunch online computer program. After determination of the chi-square value, the *P*-value approach was used to accept or reject the null hypothesis (Table5).

It has been pointed out that the F-104 was not a very forgiving aircraft. Moreover, it has been suggested that a correlation exists between pilot experience in flying hours and accident rates (Knecht, 2012, 2013; Panitzki, 1966). The chi-square independence test is suitable to examine correlations between flying hours and causation type.

The assumptions for a chi-square independence test are the same as for the chi-square goodness-of-fit test: (a) all expected frequencies are 1 or greater; (b) at most 20% of the expected frequencies are less than 5; and (c) simple random sample (Weiss, 2016).

Assumption (a) has again been fulfilled. Assumption (b) could not be fulfilled, solely due to the reduced sample size (Tables 7 and 8). Assumption (c) was once more not fulfilled, due to the quasi-experimental research design. Still, the method will be performed and the reduced statistical power will be addressed.

Following the HFACS coding, Table 7 was constructed, which depicts the observed frequencies of error types in relation to flying hours on aircraft type. For the flying hours on type, 500-hour intervals were chosen. Table 7 was developed into a contingency table (Table 8) which depicts both observed and expected frequencies for each error type in relation to flying hours on type.

The chi-square independence test calculations were performed with the StatCrunch online computer program. After determination of the chi-square value, the *P*-value approach was used to accept or reject the null hypothesis (Table 9). H2 up to and including H5 can be tested with descriptive statistics (Table 6) and numerical comparison (Weiss, 2016).

Shappell and Wiegmann (2001, 2003) as well as Ergai (2013) have described in detail why the HFACS framework can be regarded as a valid methodology for post-accident analyses. They have argued that content validity, face validity, and construct validity are essential for any taxonomy to be usable and that those are clearly present in HFACS. Content validity refers to a method's comprehensiveness and reliability. Both have already been sufficiently addressed and it can therefore be inferred that the HFACS method has a high content validity.

Face validity refers to whether a method really measures what it is meant to measure and goes hand in hand with content validity. Shappell and Wiegmann (2001, 2003) as well as Ergai (2013) have delineated that research on HFACS as a methodology has demonstrated the framework's face validity. Construct validity relates to a taxonomy's ability to close the gap between theory and practice. In the case of HFACS as human error taxonomy the question must be answered as to whether this framework only describes errors and accidents, or whether it aids in identifying the underlying causes. Shappell and Wiegmann (2001, 2003) as well as Ergai (2013) have outlined that research on HFACS as a methodology has supported the existence of construct validity of the framework. Per definition quasi-experimental designs have a lower internal validity than true experiments (Bradley, 2018). To offset the reduced internal validity of quasi-experiments, a large sample size was meant to increase the research design's accuracy. The desired sample size for high internal validity was derived *a priori* with the G*Power computer program (see Statistical power determination section).

In addition to the statistical power prediction, the research design and methodology follow established scientific standards. Hence, this research is expected to have a high internal validity (Michael, n.d.).

Statistical power describes the probability that research will detect a certain effect, if this effect exists in the first place. This effect is usually expressed as large (i.e., 0.5), medium (i.e., 0.3), or small (i.e., 0.1). High statistical power is desired since it means statistically significant results, whereas low statistical power may mean inconclusive results. A statistical power level of 0.8 is commonly used for meaningful results (Brownlee, 2018; Power Analysis, n.d.).

Thus, a power analysis should be performed before conducting an experiment or quasi-experiment to calculate the desired sample size to achieve meaningful results (Bower, 2008; Brownlee, 2018; Power Analysis, n.d.). To this end three chi-square *a priori* tests were conducted.

In order to achieve a very high statistical power of 0.95 (0.80 being commonly used) at a significance level of $\alpha = 0.05$ with a *large* effect size ($w = 0.5$) and DF = 3 (for a 4×2 contingency table), a total sample size (i.e., test group plus control group) of 69 is required (Brownlee,

2018; Faul et al., 2007; G*Power, 2017; Mayr et al., 2007). Hence, the planned test group sample of 71 was appropriate (total sample size $n = 143$).

In order to achieve the commonly accepted statistical power of 0.8 at a significance level of $\alpha = 0.05$ with a *medium* effect size ($w = 0.3$) and DF = 3 (for a 4 × 2 contingency table), a total sample size (i.e., test group plus control group) of 122 is required (Brownlee, 2018; Faul et al., 2007; G*Power, 2017; Mayr et al., 2007). Hence, the planned test group sample of 71 was again appropriate (total sample size $n = 143$).

The HFACS coding led to a reduced test group sample of only 34 accidents (i.e., total sample size $n = 106$). Hence, another *a priori* test was conducted. In order to achieve a slightly lower statistical power level than commonly used (i.e., 0.73) at a significance level of $\alpha = 0.05$ with a medium effect size ($w = 0.3$) and DF = 3 (for a 4 × 2 contingency table), a total sample size (i.e., test group plus control group) of $n = 104$ is required (Brownlee, 2018; Faul et al., 2007; G*Power, 2017; Mayr et al., 2007). Hence, the predicted statistical power of the chi-square hypotheses tests will be slightly lower than commonly used.

External validity means that a generalization of results is possible. Among other prerequisites, a scientific research method is required, as well as high internal validity, a representative test group, and sufficient statistical power (Michael, n.d.). In the presented research the scientific method, internal validity, and the test group have been sufficiently addressed to enable generalization. Thus, the greatest challenge to the presented research's external validity is the statistical power of the chi-square analyses. Therefore, a *post hoc* analysis was conducted with the G*Power computer program.

Statistical power determination

For the chi-square goodness-of-fit test's *post hoc* analysis the expected and observed frequencies of Table 4 were used. G*Power calculated an existing medium effect size of $w = 0.307$ at a significance level of $\alpha = 0.05$ and DF = 3 (for a 4 × 2 contingency table). For a total sample size of $n = 106$ (i.e., test group plus control group) a statistical power ($1 - \beta$) of 0.76 was calculated (Faul et al., 2007; G*Power, 2017; Mayr et al., 2007).

Hence, for an observed medium effect size, the computed statistical power is slightly lower than commonly accepted, but higher than predicted (Brownlee, 2018). For the originally planned total sample size of 143 the statistical power would have been 0.88 with otherwise unchanged parameters. Although the reduction of the test group sample had a negative effect on the presented research's statistical power, it still produced meaningful results with an effect size of $w = 0.307$ and a power of 0.76.

G*Power cannot be used to calculate the statistical power of chi-square independence tests. But since two out of three assumptions could not be fulfilled the results will be regarded as chi-square suspect and addressed in the results section.

Results

H1 Hypothesis Test (Inferential Statistics)

It was hypothesized that the relative distribution of Errors and Violations contributing to German F-104 losses does not differ significantly from USAF accidents. This hypothesis was tested with a chi-square goodness-of-fit test. Tables 3 and 4 show the HFACS Level-1 coding results for the test and control groups.

Table 4 shows the chi-square goodness-of-fit contingency table, which is used to test H1. Table 3 shows the differences between German F-104 expected and observed relative frequencies, which are used to test H2 to H5. The single most striking result is the large number of Skill-Based Errors in German F-104 accidents (i.e., 55.56%).

To accept or reject $H1_0$, the chi-square goodness-of-fit test must be continued either with the critical value approach or with the *P*-value approach (Weiss, 2016). Since StatCrunch provides the *P*-value, the latter approach is chosen.

Table 5 shows a *P*-value of 0.0242. It must be checked whether $P \leq \alpha$ (α being the significance level of 0.05) to reject $H1_0$; if $P > \alpha$, $H1_0$ should not be rejected.

In this case *P* (0.0242) is smaller than α (0.05). Hence, $H1_0$ should be rejected and $H1_A$ should be favored.

This means that at the 5% significance level there is sufficient evidence that the relative distribution of Errors and Violations contributing to German F-104 losses differs significantly from USAF accidents.

H2 Hypothesis Test (Descriptive Statistics)

It was hypothesized that the number of Decision Errors contributing to German F-104 losses does not differ significantly from USAF accidents.

Decision Errors contributed to 44% of USAF accidents (Table 6). Therefore, in order to favor $H2_0$ and reject $H2_A$, Decision Errors must have contributed to German F-104 accidents in a span of 41.8% to 46.2%. However, Decision Errors only contributed to 35.29% of German F-104 accidents. Hence, $H2_0$ should be rejected and $H2_A$ should be favored.

This means that at the 5% significance level there is sufficient evidence that the number of Decision Errors contributing to German F-104 losses differs significantly from USAF accidents.

Table 3
HFACS coding results.

Accident causation	USAF: observed relative frequency 1991–1997	German F-104: expected relative frequency 1978–1986	German F-104: observed relative frequency 1978–1986
Decision Errors	44	44	35.29
Skill-Based Errors	60	60	73.53
Perceptual Errors	33	33	17.65
Violations	8	8	5.88

Note. The results do not add up to 100%, since accidents usually have multiple causations.

Table 4
Chi-square goodness-of-fit contingency table.

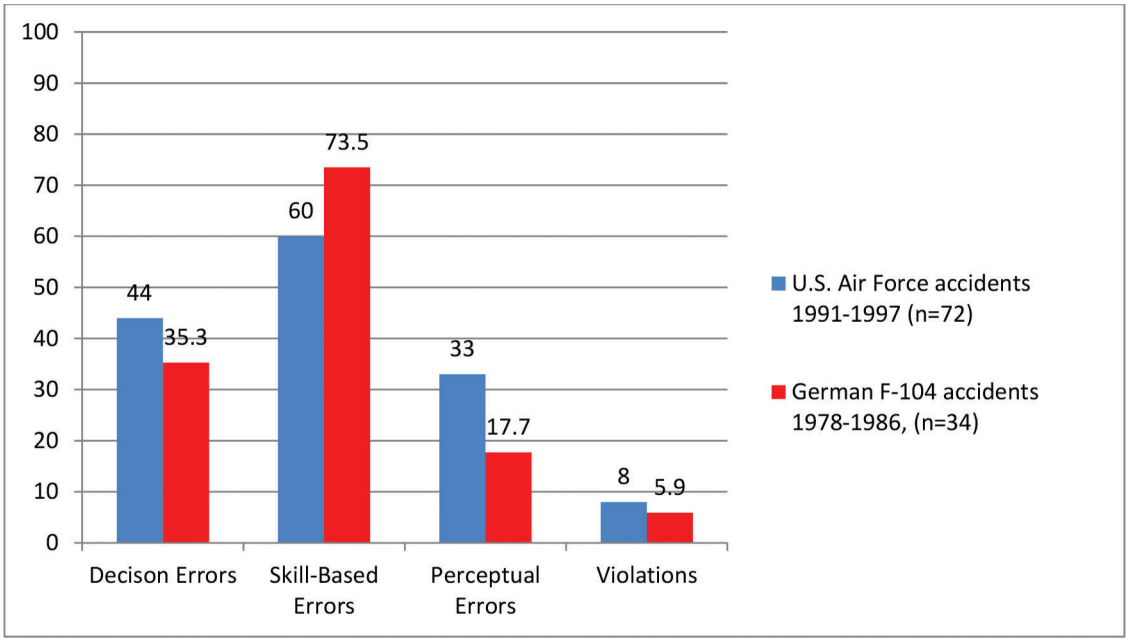
Accident causation	USAF: observed relative frequency 1991–1997	German F-104: expected relative frequency 1978–1986	German F-104: observed relative frequency 1978–1986
Decision Errors	30.34	30.34	26.66
Skill-Based Errors	41.38	41.38	55.56
Perceptual Errors	22.76	22.76	13.34
Violations	5.52	5.52	4.44
Total	100	100	100

Note. Results after factorization (i.e., 0.689 for USAF and 0.7556 for F-104) to achieve totals of 100. Numerical relations of Table 3 are unchanged.

Table 5
Chi-square goodness-of-fit results.

N	DF	Chi-square	*P*-value
100	3	9.415615	0.0242

Table 6
Histogram of HFACS Level-1 coding results (observed relative frequency according Table 3).



Note. The results do not add up to 100%, since accidents usually have multiple causations.

H3 Hypothesis Test (Descriptive Statistics)

It was hypothesized that the number of Skill-Based Errors contributing to German F-104 losses does not differ significantly from USAF accidents.

Skill-Based Errors contributed to 60% of USAF accidents (Table 6). Therefore, in order to favor $H3_0$ and reject $H3_A$, Skill-Based Errors must have contributed to German F-104 accidents in a span of 57% to 63%. However, Skill-Based Errors contributed to 73.5% of German F-104 accidents. Hence, $H3_0$ should be rejected and $H3_A$ should be favored.

This means that at the 5% significance level there is sufficient evidence that the number of Skill-Based Errors contributing to German F-104 losses differs significantly from USAF accidents.

H4 Hypothesis Test (Descriptive Statistics)

It was hypothesized that the number of Perceptual Errors contributing to German F-104 losses does not differ significantly from USAF accidents.

Perceptual Errors contributed to 33% of USAF accidents (Table 6). Therefore, in order to favor $H4_0$ and reject $H4_A$, Perceptual Errors must have contributed to German F-104 accidents in a span of 31.35% to 34.65%. However, Perceptual Errors only contributed to 17.65% of German F-104 accidents. Hence, $H4_0$ should be rejected and $H4_A$ should be favored.

This means that at the 5% significance level there is sufficient evidence that the number of Perceptual Errors contributing to German F-104 losses differs significantly from USAF accidents.

H5 Hypothesis Test (Descriptive Statistics)

It was hypothesized that the number of Violations (Routine and Exceptional) contributing to German F-104 losses does not differ significantly from USAF accidents.

Violations contributed to 8% of USAF accidents (Table 6). Therefore, in order to favor $H5_0$ and reject $H5_A$, Violations must have contributed to German F-104 accidents in a span of 7.6% to 8.4%. However, Violations only contributed to 5.88% of German F-104 accidents. Hence, $H5_0$ should be rejected and $H5_A$ should be favored.

This means that at the 5% significance level there is sufficient evidence that the number of Violations contributing to German F-104 losses differs significantly from USAF accidents.

H6 Hypothesis Test (Inferential Statistics)

It was hypothesized that no strong association exists between error type and pilot flying hours on F-104.

This hypothesis was tested with a chi-square independence test. Tables 7 and 8 show the HFACS Level-1 coding results (excluding violations) and their relative distribution in 500-hour intervals.

To accept or reject $H6_0$, the chi-square independence test must be continued either with the critical value approach or with the *P*-value approach (Weiss, 2016). Since StatCrunch provides the *P*-value, the latter approach is chosen.

Table 9 shows a *P*-value of 0.574. It must be checked whether $P \leq \alpha$ (α being the significance level of 0.05) to reject $H6_0$; if $P > \alpha$, $H6_0$ should not be rejected.

In this case *P* (0.574) is larger than α (0.05). Hence, $H6_0$ should be favored while $H6_A$ should be rejected.

This means that at the 5% significance level there is no sufficient evidence for an association between error type and pilot flying hours on F-104. However, H6 should remain undetermined, since two out of three assumptions for a chi-square independence test could not be fulfilled. This will be addressed in the discussion section.

Table 7
Chi-square independence test (observed frequencies).

Accident causation	1–500 flying hours on F-104	501–1000 flying hours on F-104	>1000 flying hours on F-104	Total
Decision Errors	2	5	5	12
Skill-Based Errors	9	10	6	25
Perceptual Errors	1	2	3	6
Total	12	17	14	43

Table 8
Chi-square independence test contingency table (observed and expected frequencies).

Accident causation	1–500 flying hours on F-104	501–1000 flying hours on F-104	>1000 flying hours on F-104	Total
Decision Errors	2 (3.35)	5 (4.74)	5 (3.91)	12
Skill-Based Errors	9 (6.98)	10 (9.88)	6 (8.14)	25
Perceptual Errors	1 (1.67)	2 (2.37)	3 (1.95)	6
Total	12	17	14	43

Note. Expected frequencies in parentheses.

Other Results

This research's *first unexpected result* is the large number of accidents without any unsafe act whatsoever. From the original sample of 71 accidents, two had to be withdrawn due to unknown accident causations. This left a sample of 69 accidents. Of those 69 accidents, 17 aircraft crashed after their engine failed for technical reasons. In another 11 cases the engine failed after a bird strike; additionally, seven aircraft crashed due to technical issues, such as controllability problems, stuck fuel, or gear malfunctions (Table 10). Hence, 35 (i.e., 50.72%) of these 69 accidents occurred without any unsafe act whatsoever (HFACS Level-2). This means that human factor causes could only be attributed to 49.28% of the F-104 sample. Therefore, the German F-104 accidents *are* different from other military aircraft accidents, to begin with.

Table 9
Chi-square independence test results.

Statistic	DF	Value	*P*-value
Chi-square	4	2.9040056	0.574

Table 10
German F-104 accidents without unsafe act.

Causation	Rate ($N = 69$)
Engine failure	24.64% ($n = 17$)
Engine failure after bird strike	15.94% ($n = 11$)
Other technical failure	10.14% ($n = 7$)
Total	50.72% ($n = 35$)

Note. From the sample of $N = 71$, two accidents had to be excluded due to unknown accident causation.

Moreover, 24.64% of the sample crashed due engine failures, while another 10.14% crashed for other technical reasons. Hence, the *second unexpected result* was that more than a third of the sample's accidents (i.e., 34.78%) occurred exclusively due the Technological Environment (HFACS Level-2).

Another 15.94% were attributed to engine failures following a bird strike, which means due to the Physical Environment (HFACS Level-2) in which the aircraft was operated. Bird strikes are largely beyond an organization's or pilot's control. Hence, 15.94% of the German Starfighter accidents can be regarded as "the price of doing business" with a single-engine turbojet fighter in the low-altitude environment that Naval and Air Force strike aircraft populated at the time.

When looking at engine failures in isolation, 17 aircraft crashed after their engine failed for technical reasons (i.e., 24.64%) and another 11 after engine failure following a bird strike (i.e., 15.94%). Thus, 28 aircraft crashed after their engine failed, which constitutes 40.58% of the sample. This is the *third unexpected result* and this elevated level of engine failures needs to be addressed later on.

Summary

In five out of six cases the alternative hypothesis was supported. The chi-square goodness-of-fit test showed that the relative distribution of Errors and Violations in German F-104 accidents differs significantly from USAF accidents ($H1_A$). Moreover, the USAF accidents contained significantly more Decision Errors, significantly fewer Skill-Based Errors, significantly more Perceptual Errors, and significantly more Violations. Thus $H2_A$ to $H5_A$ were favored over $H2_0$ to $H5_0$. Lastly, it was decided that H6 should remain undetermined, since two out of three assumptions for a chi-square independence test could not be fulfilled. The latter was a direct result of the test group's necessary reduction from 71 to only 34 cases: in addition to two undetermined accident causes, 35 accidents (i.e., 50.72%) occurred without an unsafe act whatsoever, but due to the Physical Environment and/or the Technological Environment (both Level-2: Preconditions for Unsafe Acts).

Discussion

This project's research question was to find out whether the German F-104s crashed for different reasons from those for *other* military aircraft. Some results were most unexpected and, consequently, six different aspects need to be addressed in this discussion: (a) the number of German F-104 accidents without unsafe act; (b) the German F-104 accident rate in relation to that of its predecessor and in relation to USAF macro trends; (c) the German F-104 accident rate in relation to USAF F-104 accidents and other co-era USAF fighters; (d) engine-related German F-104 accidents in relation to engine-related USAF accidents; (e) the HFACS Level-1 distribution in German F-104 accidents; and (f) other findings.

It was expected that the Starfighter's safety record would be much better than its reputation suggests and that it did not crash for reasons different from those for other military aircraft. It was also expected the vast number of accidents would be attributed to human error, and would hence be in line with contemporary research (Dekker, 2002; Helmreich & Foushee, 2010; Shappell & Wiegmann, 2001).

As pointed out in the results, 50.72% of the accidents occurred without unsafe act but due to the technological and physical environment. Hence, not only did the German F-104 crash for reasons different from those for other military aircraft (H1 to H5), but it also crashed significantly more often due to technological shortcomings and the environment in which it was operated (Table 10).

It has been suggested that a different view on the German Starfighter losses is possible. The Starfighter's

predecessor in service with Germany's Fighter Bomber Wings was the Republic's F-84. The German Air Force acquired 558 F-84s of different types, of which 202 were lost (Reis, 2012). This constitutes an *attrition rate* of 36.2% for the F-84, in contrast to 31.88% for the F-104 (i.e., 916 acquired and 292 lost). Moreover, Reis (2012) and Siano (2016) have pointed out that the Starfighter's annual *accident rate* was also slightly lower than the F-84's rate.

The argument goes that since no public outcry occurred in the case of the F-84, there should not have been one in the case of the F-104 either. And after all, other F-104 users had even higher loss rates (Table 1)—hence, the Starfighter losses could be seen as less dramatic than public opinion has it.

While at first glance the previously mentioned arguments appear convincing, three aspects deserve attention. (a) It is misleading that Schlieper (1995) and Kropf (2002) express the F-104's accident rate in accidents per 10,000 flight hours; accident rates are usually expressed in accidents per 100,000 flight hours. While two accidents per 10,000 flight hours appears to be reasonably low, 20 accidents per 100,000 flight hours is not a low rate at all. Besides, for most German units this accident rate meant an average loss of two aircraft per year, and one fatality every 1.5 years. (b) From 1950 onwards the USAF's overall accident rate has been steadily declining and after 1965 it has been *constantly below five accidents per 100,000* flying hours (Kitfield, 1996; Shappell & Wiegmann, 2003; USAF, 2000). A comparison with other USAF fighter types follows in subsequent paragraphs. Hence, although the German F-104's accident and attrition rates were slightly lower than those of its legacy predecessor, this accident rate was still three to four times higher than the USAF standard from 1965 onwards. (c) Public outcry or not, the F-84's attrition rate should by no means be regarded as normal or acceptable (see below).

Lyons and Nace (2007) have reviewed and analyzed 25 historical USAF airframes and compared their accident rates (Table 1). For the USAF's F-104 they found a final cumulative crash rate of 306.3 crashes per million flight hours—i.e., 30.63 accidents per 100,000 flying hours during its 28 years in the USAF inventory.

So, on the one hand, the German F-104 accident rate was finally *distinctly lower* than the USAF's F-104 accident rate (15 to 20 accidents versus 30.63 accidents per 100,000 flying hours); but on the other hand, both the USAF F-104's and the German F-104's accident rates were *distinctly higher* than the standard USAF accident rate (five or less accidents per 100,000 flying hours; see previous discussion).

Besides, even this "reduced" German F-104 accident rate would still grant it a top ranking in comparison to co-era USAF fighters (Table 1). Only the USAF F-104's accident rate is even higher (30.36 accidents per 100,000 flight hours), followed by F-100 (21.22), F-105 (17.83), F-101 (14.65), F-102 (13.69), F-106 (9.47), and F-5 (8.82).

This means that both the German and the USAF F-104 accident rates are unsurpassed by any co-era or succeeding type in the USAF inventory. Hence, it is very reasonable to infer that the F-104 as a design was much more accident prone than other contemporary U.S. types.

It has been pointed out that 40.58% of the reviewed German F-104 accidents occurred due to engine failures (i.e., 24.64% solely for technical reasons, 15.94% after a bird strike; Table 10). This raises the question as to whether the F-104's J-79 engine should be regarded as unreliable and/or susceptible.

While the number of engine-related German F-104 accidents is surprising, the USAF F-104 accidents indicate the same trend (Table 2). To this date the F-104 ranks again *by far* on top of the list of USAF engine-related accidents—outranking each competitor by a factor of two or more. Moreover, the USAF F-104's engine-related accident rate of 9.48 accidents means it crashed twice as often as the USAF average for engine issues alone.

When recognizing the USAF F-104's total accident rate of 30.36 accidents per 100,000 flight hours, it is striking that for both German and U.S. F-104s about a third of their accidents occurred due to engine failures.

Thus, it is reasonable to regard the J-79 engine as a weak link in the F-104's safety record.

It was expected that in contrast to public belief, the German F-104s crashed for no other reasons different from those for contemporary fighters. Three unexpected results have already been addressed, indicating an underrepresentation of human factor causations.

The *fourth unexpected result* is the rejection of $H1_0$ to $H5_0$. In all cases the alternate hypothesis had to be favored. This means: not only did the German F-104s crash for different reasons from those for other aircraft, it also means that if human factor causations were involved, those also showed a significantly different distribution in relation to the USAF control group.

There is no simple explanation for the different distribution in the HFACS coding (Table 6). The results show significantly fewer Decision Errors (i.e., 35.3% versus 44%), significantly fewer Perceptual Errors (i.e., 17.7% versus 33%), and significantly fewer Violations (i.e., 5.9% versus 8%). This may be an indicator for well trained and disciplined pilots, but it appears impossible to pinpoint the exact causation. After all, it was expected that the relative distribution was no other than in the control group.

On the other hand, the results show significantly more Skill-Based Errors than in the control group (i.e., 73.5% versus 60%). This result supports Rall's (2004) and Vogler's (1986) judgment of a complex design, which took lots of practice to master and which permitted only small error margins.

Unfortunately, the large number of accidents without unsafe acts led to a sample size reduction. This implied that two out of three assumptions for the chi-square independence test could not be fulfilled. The lack of randomization is inherent to a quasi-experimental design and was meant to be overcome by a large sample size. This was denied by the sample size reduction. Additionally, it implied that the third assumption (i.e., no less than 20% of expected frequencies should be below 5) could not be fulfilled either. Therefore, the results of the independence test will be regarded as chi-square suspect, and hence H6 must remain undetermined. Although it very likely exists, there is presently no sufficient evidence for any correlation between flying hours on type and error type.

Moreover, the accident investigations' quality and thoroughness surprised and impressed the author. On average the accident files contained 250 pages, 534 pages being the maximum. During this research, even earlier F-104 accidents from the 1960s were reviewed and showed the same thoroughness. Even the oldest reports showed professionalism of the highest standards and a stunning attention to detail. The final reports' intonations were generally fair and usually free of pilot blaming. Besides, not a single case of alcohol or substance abuse was found.

Another issue which became evident when reviewing the files was the excessive time span from identifying a problem to fixing it. The following is but one of many examples. The official report of GenFlSichhBw on the second F-104 accident in September 1961 recommended the development and installation of a flight data recorder. It took until 1974 until the LEADS 200 device (Leigh Electronic Airborne Data System) was installed into the first F-104s (BMVg, 1985; GenFlSichhBw, 1961). Until the end of its service life only 50 F-104s were equipped with the LEADS 200 device, and in those aircraft the LEADS 200 was deactivated repeatedly due to malfunctions (Das Flugdatenregistriersystem, n.d.).

Conclusion

Approximately 15–20% of the German F-104 accidents can be called "the price of doing business" with such an aircraft in the European environment. This includes the operation of a very demanding single-seat, single-engine turbojet fighter close to the ground—and in a bird-rich environment. Still, both German and U.S. F-104s had significantly elevated accident rates in comparison to other co-era types. The number of technical causations was grossly overrepresented in German F-104 accidents, while the number of human factor causations was underrepresented. Moreover, the engine was found to be a weak part in the design, which holds a negative record in the USAF even to this date.

Hence, it can be concluded that the F-104 *did* crash for different reasons from those for other military aircraft, and it also crashed for other human factor reasons. Overall, the Starfighter was more accident-prone than its co-era types.

List of Abbreviations

Abbreviation	English term	German term
AOA	Angle of attack	
BMVg	German Ministry of Defense	Bundesministerium der Verteidigung
GenFlSichhBw	Directorate of Aviation Safety, Federal Armed Forces	General Flugsicherheit der Bundeswehr
CRM	Crew resource management	
HFACS	Human Factors Analysis and Classification System	
LEADS 200	Leigh Electronic Airborne Data System 200	
NATO	North Atlantic Treaty Organization	
NOTECHS	Non-technical skills	
MOD	Ministry of Defense	
USAF	United States Air Force	
WTD	Military Flight Test Installation	Wehrtechnische Dienststelle

References

Bower, K. (2008). What is statistical power? Retrieved from https://www.youtube.com/watch?v=z-D6VmG7aLA

Bowman, M. (2000). *Lockheed F-104 Starfighter*. Wiltshire, UK: Crowood Press.

Bradley, K. (2018). Quasi-experiment advantages & disadvantages. Retrieved from https://classroom.synonym.com/quasiexperiment-advantages-disadvantages-8614272.html

Brownlee, J. (2018). A gentle introduction to statistical power and power analysis in Python. Retrieved from https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/

Bundesministerium der Verteidigung (Ed.). (1985). Flight manual GAF series F/TF-104G.

Das Flugdatenregistriersystem—LEADS 200 [Flight Data Recorder LEADS 200]. (n.d.). Retrieved from http://www.rolfferch.de/F104G/html/sondereinbauten.html

Dekker, S. (2002). *The field guide to human error investigations*. New York, NY: Ashgate Publishing.

Ein schöner Tod—fürs Vaterland? [A beautiful death – for the fatherland?] (1982). *Der Spiegel*, *35*, 94–97. Retrieved from http://www.spiegel.de/spiegel/print/d-14349673.html

Ergai, A. O. (2013). *Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability* [Doctoral dissertation, Ann Arbor, MI]. ProQuest LLC.

Ergai, A., Cohen, T., Sharp, J. Wiegmann, D, Gramopadhye, A., & Shappell, S. (2016). Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability. *Safety Science*, *82*, 393–398.

F-104. Holder of the absolute records for speed and altitude. (1958). *Flight,*30 May*1958*, 739-43. Retrieved from http://www.916-starfighter.de/Record%20Holder%20F-104_Flight_30May1958.pdf

Faul, F., Erdfelder, E., Lange, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fischbach, G. (1998). *916 deutsche F-104 Starfighter. Ihre Bau- und Lebensgeschichten* [916 German Starfighters. Their production- and life stories]. Holzkirchen, Germany: self-publishing.

Flin, R., Martin, L., Goeters, K. M., Hörmann, H. J. Amalberti, R, Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilot's CRM skills. *Human Factors and Aerospace Safety*, *3*(2), 95–117.

G*Power 3.1 manual. (2017). Retrieved from http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf

General Flugsicherheit der Bundeswehr. (1961). Accident file 61106.

Helmreich, R. L., & Foushee, H. C. (2010). Why CRM? Empirical and theoretical bases of human factors training. In Kanki, B. Helmreich R. & Anca J. (Eds.), *Crew resource management* (pp. 3–58). San Diego, CA: Elsevier.

Hooper, B. J., & O'Hare, D. P. A. (2013). Exploring human error in military aviation flight safety events using post-incident classification systems. *Aviation, Space, and Environmental Medicine*, *84*(7), 1–11.

Kauf von Schrott [Purchase of scrap metal]. (1969). *Der Spiegel*, *43*, 107–111. Retrieved from http://www.spiegel.de/spiegel/print/d-45520594.html

Kitfield, J. (1996, June). Flying safety: The real story. *Air Force Magazine*. Retrieved from http://www.airforcemag.com/MagazineArchive/Documents/1996/June%201996/0696safety.pdf

Knecht, W. R. (2012). *Predicting general aviation accident frequency from pilot total flight hours*. Washington, DC: FAA.

Knecht, W. R. (2013). The "killing zone" revisited: Serial nonlinearities predict general aviation accident rates from pilot total flight hours. *Accident Analysis and Prevention*, *60*, 50–56.

Kropf, K. (2002). *German Starfighters. The F-104 in German air force and naval air service*. Hinckley, UK: Midland Publishing.

Lemke, B. (2006). Konzeption und Aufbau der Luftwaffe [The German Air Force's concept and organizational structure]. In Lemke, B. Krüger, D. Rebhan H. & Schmidt W. (Eds.), *Die Luftwaffe 1950 bis 1970. Konzeption, Aufbau, Integration* [The German Air Force 1950 to 1970. Its concept, structure and organizational integration] (pp. 71–484). Munich, Germany: Oldenbourg Verlag.

Lockheed (Ed.). (1960). F-104G flight manual.

Loy, H. (2011). *Jahre des Donners. Mein Leben mit dem Starfighter* [Years of thunder. My life with the Starfighter]. Rosenheim, Germany: Rosenheimer Verlagshaus GmbH & Co. KG.

Lyons, T. J., & Nace, W. (2007). Aircraft crash rates and cumulative hours: USAF data for 25 airframes, 1950–2006. *Aviation, Space, and Environmental Medicine*, *78*(10), 923–925.

Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 51–59.

Michael, R. S. (n.d.). Threats to internal & external validity. Retrieved from http://www.indiana.edu/~educy520/sec5982/week_9/520in_ex_validity.pdf

Neu, M. (2004). Die Starfighter-Beschaffung—ein politischer Skandal? [The Starfighter acquisition—a political scandal?]. In Bellers J. & Königsberg M. (Eds.), *Skandal oder medienrummel* [Scandal or media hype] (pp. 65–120). Münster, Germany: LIT Verlag.

North Atlantic Treaty Organization (1957). MC 14/2 (Revised). Overall strategic concept for the defense of the North Atlantic Treaty Organization area. Retrieved from https://www.nato.int/docu/stratdoc/eng/a570523a.pdf

Panitzki, W. (1966). Verteidigungsausschuss, Protokoll 5. Wahlperiode, 4. Sitzung [Defense Committee, protocol, 5th election period, session 4], January 14th, 1966. In Siano, C. *Die Luftwaffe und der Starfighter* [The German Air Force and the Starfighter]. (p.277).

Power Analysis, Statistical Significance, & Effect Size. (n.d.). Retrieved from http://meera.snre.umich.edu/power-analysis-statistical-significance-effect-size

Rall, G. (2004). *Mein Flugbuch. Erinnerungen 1938–2004* [My flight log. Memories 1938–2004]. Moosburg, Germany: NeunundzwanzigSechs Verlag.

Reason, J. (1990). *Human error*. New York, NY: Cambridge University Press.

Reason, J. (2000). Human error: models and management. *British Medical Journal*, *320*, 768–770. doi: http://10.1136/bmj.320.7237.768

Reaves, G. L. (1961). *The F-104 Starfighter. Test pilot's notebook*. Burbank, CA: Lockheed Aircraft Corporation. Retrieved from http://www.916-starfighter.de/Snake%20Reaves_Test%20Pilot%20Notebook%20F-104.PDF

Rebhan, H. (2006). Aufbau und Organisation der Luftwaffe 1955 bis 1971 [Structure and organization of the German Air Force 1955 to 1971]. In Lemke, B. Krüger, D. Rebhan H. & Schmidt W. (Eds.), *Die Luftwaffe 1950 bis 1970. Konzeption, Aufbau, Integration* [The German Air Force 1950 to 1970. Its concept, structure and organizational integration] (pp. 3–37). Munich, Germany: Oldenbourg Verlag.

Reis, S. (2012). Das Krisenmanagement der Luftwaffe: Die Bewältigung der Starfighter-Krise [German Air Force crisis management. Overcoming the Starfighter crisis]. In E. Birk, H. Möllers & W. Schmidt (Eds.), *Die Luftwaffe zwischen Politik und Technik* [The German Air Force between politics and technology] (pp. 88–107). Berlin, Germany: Carola Hartmann Miles-Verlag.

Schlieper, A. (1995). Die Wechselwirkung Taktik—Technik—Mensch. Die Einführung des Flugzeugs F-104G in die deutsche Luftwaffe und die Starfighterkrise von 1965/66 [Interaction of tactics—technology—and human being. The introduction of the F-104G in the German Air Force and the Starfighter Crisis of 1965/66]. In Thoss B. & Schmidt W. (Eds.), *Vom Kalten Krieg zur deutschen Einheit* [From Cold War to the

German re-unification] (pp. 551–583). Munich, Germany: Oldenbourg Verlag.

Shappell, S. A., & Wiegmann, D. A. (2000). The Human Factors Analysis and Classification System—HFACS. Retrieved from https://commons.erau.edu/cgi/viewcontent.cgi?article=1777&context=publication

Shappell, S. A., & Wiegmann, D. A. (2001). Applying reason: The human factors analysis and classification system (HFACS). *Human Factors and Aerospace Safety*, *1*(1), 59–86.

Shappell, S. A., & Wiegmann, D. A. (2003). *A human error approach to aviation accident investigation. The Human Factors Analysis and Classification System*. Burlington, VT: Ashgate Publishing.

Shappell, S. A., & Wiegmann, D. A. (2004). HFACS analysis of military and civilian aviation accidents: A North American comparison. International Society of Air Safety Investigators, *Proceedings of the 35th Annual International Seminar*, *2004* (pp. 135–140). Retrieved from https://www.isasi.org/Documents/library/Seminar-Proceedings/Proceedings-2004.pdf

Siano, C. (2016). *Die Luftwaffe und der Starfighter* [The German Air Force and the Starfighter] [Doctoral dissertation]. Berlin, Germany: Carola Hartmann Miles-Verlag.

Stiller, G. (1981, November 29). Yogi-bär, wollen wir aussteigen? [Yogi-bear, shall we eject?]. *Bild am Sonntag*, 34.

U.S. Air Force (Ed.). (1960). Flight manual F-104D. USAF series aircraft. Retrieved from http://www.916-starfighter.de/F-104D_Flight%20Manual.pdf

U.S. Air Force (Ed.). (2000). Air Force system safety handbook. Retrieved from https://www.system-safety.org/Documents/AF_System-Safety-HNDBK.pdf

U.S. Air Force (Ed.). (2015a). USAF engine-related fighter/attack Class A flight mishap rates for single engine aircraft. Retrieved from https://www.safety.af.mil/Portals/71/documents/Aviation/Engine%20Statistics/USAF%20Single%20Engine.pdf

U.S. Air Force (Ed.). (2015b). USAF engine-related fighter/attack Class A flight mishap rates for twin engine aircraft. Retrieved from https://www.safety.af.mil/Portals/71/documents/Aviation/Engine%20Statistics/USAF%20Twin%20Engine.pdf

Vogler, P. (1986). Open letter to the F-104. In Kropf K. (Ed.), *German Starfighters. The F-104 in German air force and naval air service* (pp. 8–10).

Weiss, N. A. (2016). *Introductory statistics* (10th ed.). New York, NY: Pearson Education.

Wettig, G. (1995). Von der Entmilitarisierung zur Aufrüstung in beiden Teilen Deutschlands 1945–1952 [From demilitarization to re-armament in both parts of Germany]. In Thoss B. & Schmidt W. (Eds.), *Vom Kalten Krieg zur deutschen Einheit* [From Cold War to the German re-unification] (pp. 3–37). Munich, Germany: Oldenbourg Verlag.