Letter to the Editor re: Evaluating the Reliability of the Human Factors Analysis Classification System

Dear Editor:

The article "Evaluating the reliability of the Human Factors Analysis and Classification System," by Cohen, Shappell, and Wiegmann,[1] misrepresents three of the articles[3–5] used in their systematic review. The authors' listing of "Reliable" for interrater reliability (IRR) in Table III does not give a fair representation of what these three authors actually found and stated in their studies. This misrepresentation seems to be based on an overreliance on the overall IRR value. In these three articles, the authors state that although the overall IRR was acceptable, the reliability of the causal factors was not acceptable, and that "the acceptable overall reliability can be attributed to the high reliability in rejecting nanocodes that clearly did not apply to the mishaps." Therefore, five of the six articles which specifically examined IRR found HFACS to be unreliable.

If the DoD-HFACS articles are discounted due to increased granularity in the data, three articles are left; however, as stated in Cohen et al.'s article, two of those articles claim the rating is unreliable. The remaining study, based on only two raters, found the rating reliable. Although 14 studies were included in this review, 8 did not specifically examine reliability. Furthermore, the only studies that included data from more than 10 raters were the 3 that were misrepresented.

Given that HFACS has 19 causal categories and DoD-HFACS has 149 nanocodes, it seems reasonable that IRR should be analyzed separately. To that end, all three studies (plus Hughes' unpublished study[2]) that have tested DoD-HFACS specifically for IRR have found it unreliable. Two of the three studies that have tested HFACS specifically for IRR have found it unreliable. Based upon the studies to date there are insufficient data to support the reliability of HFACS.

O'Connor[5] recommends that coding systems be evaluated for reliability and validity prior to widespread implementation. In fact, as of 1 October 2015, the U.S. Air Force no longer codes Ground Class C and D mishaps because of the unreliability of DoD-HFACS.

Bruce R. Burnham, CSP
Air Force Safety Center, Kirtland AFB, NM

ACKNOWLEDGMENTS

The views expressed herein are the view of the author and do not reflect the official policy of the Department of the Air Force, the Department of Defense, or of the U.S Government.

REFERENCES

1. Cohen TN, Wiegmann DA, Shappell SA. Evaluating the reliability of the Human Factors Analysis and Classification System. Aerosp Med Hum Perform. 2015; 86(8):728–735.
2. Hughes TG, Heupel KA, Musselman BT, Hendrickson E. Preliminary investigation of the interrater reliability of the Department of Defense Human Factors Accident and Classification System in USAF mishaps [Abstract]. Aviat Space Environ Med. 2007; 78(3):255.
3. O'Connor P. HFACS with an additional layer of granularity: validity and utility in accident analysis. Aviat Space Environ Med. 2008; 79(6):599–606.
4. O'Connor P, Walker P. Evaluation of a human factors analysis and classification system as used by simulated mishap boards. Aviat Space Environ Med. 2011; 82(1):44–48.
5. O'Connor P, Walliser J, Philips E. Evaluation of a human factors analysis and classification system used by trained raters. Aviat Space Environ Med. 2010; 81(10):957–960.

In Response:

We reviewed published studies that examined the reliability of HFACS when used as a tool for classifying human factors data associated with accidents.[1] HFACS was not typically used during the original investigation; rather, it was used post hoc to group (code) existing causal factors into various categories. The coding process generally involved two or more coders independently performing the classification task. The agreement levels among coders (inter-rater reliability) reported in these studies was the topic highlighted in our review.

Our analysis revealed a range of reported reliability levels, some acceptable, others not. We also identified several methodological issues that could account for such discrepancies,

DOI: https://doi.org/10.3357/AMHP.4555.2017

including the lack of coder training, modifications to the HFACS framework, the use of a small number of accident cases, requiring coders to perform a two-step classification task (i.e., identifying causal factors before classifying them), and the reporting of coder agreement levels when classifying variables into categories that are non-HFACS specific. We subsequently conducted a study to ensure coders were sufficiently trained on HFACS and were able to demonstrate acceptable levels of reliability.[2]

Dr. Burnham appears to disagree with our assessment of the literature, particularly as it relates to one derivative of HFACS, namely HFACS-DoD. For those who are not aware of the distinction between the two systems, the original HFACS framework is hierarchical in nature, consisting of 4 broad levels (organizational influences, unsafe supervision, preconditions for unsafe acts, and unsafe acts) with 19 major causal-factor categories distributed across these 4 levels. The HFACS-DoD framework, as used in previously published studies, consists of the same hierarchical structure with 20 major categories that are very similar to those contained in the original HFACS. However, the DoD has also added an additional 147 subcategories, called nanocodes, that are embedded within its 20 major categories.

It is not uncommon for organizations to develop nanocodes when implementing HFACS. Nanocodes are exemplars of the types of factors that each major HFACS category represents within that domain (e.g., for aviation, one nanocode under decision errors might be "pilot misdiagnosed engine failure"). Such codes are important for archiving specific details of each accident and can help with trending data or performing more fine-grained analyses within each HFACS category. Nanocodes, by their very nature, are highly specific to a particular type of operation and, therefore, differ considerably in their number and type across organizations, industries, or specialties. Nanocodes are not HFACS, but rather idiosyncratic codes developed based on the needs and background of the user.

We attempted to distinguish between reported reliability levels based on agreements among coders when classifying data across the 4 levels and 19 major HFACS categories vs. agreement among coders when classifying factors across the nanocode categories. Our goal was to disambiguate conclusions made about the reliability of the overall HFACS structure and reliability related to the selection of appropriate nanocodes.

The assertion by Dr. Burnham that we intentionally misrepresented the data pertaining to the reliability of DoD-HFACS is unfortunate. More unfortunate, however, is the statement by Dr. Burnham that the DoD has chosen to abandon the use of HFACS for Class C and D mishaps, based, at least in part, on the three cases reported by O'Conner et al.[3–5] To discard the overall HFACS framework rather than reduce the 147 nanocode categories seems imprudent. However, such consequences should serve as a caution to others who may be considering the adoption of the HFACS framework. Although changes or additions to the framework during implementation may seem appropriate, such changes can clearly impact the fidelity and usability of the framework, which, in turn, can minimize the effectiveness that it might otherwise have.

Scott A. Shappell, Ph.D.
Department of Human Factors, College of Arts and Sciences Embry Riddle Aeronautical University, Daytona Beach, FL

REFERENCES

1. Cohen T, Wiegmann D, Shappell S. Evaluating the reliability of the Human Factors Analysis and Classification System. Aerosp Med Hum Perform. 2015; 86(8):728–735.
2. Ergai A, Cohen T, Wiegmann D, Sharp J, Gramopadhye A, Shappell S. Assessment of the Human Factors Analysis and Classification System (HFACS): intra-rater and inter-rater reliability. Safety Sci. 2016; 82:393–398.
3. O'Connor P. HFACS with an additional layer of granularity: validity and utility in accident analysis. Aviat Space Environ Med. 2008; 79(6):599–606.
4. O'Connor P, Walker P. Evaluation of a human factors analysis and classification system as used by simulated mishap boards. Aviat Space Environ Med. 2011; 82(1):44–48.
5. O'Connor P, Walliser J, Philips E. Evaluation of a human factors analysis and classification system used by trained raters. Aviat Space Environ Med. 2010; 81(10):957–960.