ABSTRACT BODY:
Purpose: A recent survey by the authors explored educational scholars’ experiences with multiple institutional review board (IRB) reviews for health professions educational trials involving more than one institution. In the initial analysis, survey participants identified significant variability in several IRB decisions such as study risk determinations and review arrangements across the multiple IRBs involved. This follow-up analysis provides insight into the types of IRB protocol review disagreements across multiple IRBs.
Approach/Methods: In the summer of 2015, medical school faculty from the U.S., Canada, and the Puerto Rico were invited to participate in a survey via three regional AAMC membership listservs and direct email solicitation with a snowball sample of medical education scholars in the 4th region. One survey item elicited these participants to reflect on the types of agreement or disagreement between IRBs decisions for six categories of study procedures, including protocol wording, recruitment procedures, inclusion/exclusion criteria, informed consent procedures, risks/benefits, and data storage and security. This study was approved by the University of Georgia IRB.
Results/Outcomes: 94 survey participants indicated that they had participated in at least one multi-site study that required review by two or more IRBs in the prior five years. For those who could recall some of the details about their most recent study (n=50), 21 noted disagreement between two or more IRBs for at least one of the six categories: 7 reported on disagreement on one category; 9 reported on disagreement on two categories; 1 reported on disagreement on three categories; 3 reported on disagreement on four categories, and; 1 reported on disagreement on five categories. All other responses were “I don’t know.”
IRBs most frequently disagreed on their request for changes to protocol wording (16/48, 33.3%) followed by informed consent procedures (14/48, 29.2%), data security procedures (6/46, 13.0%), risk/benefit determinations (5/46, 10.9%), recruitment procedures (4/49, 8.2%), and inclusion/exclusion criteria (4/50, 8.0%).
Only 8.3% participants who experienced disagreement between IRBs (3/36) believed that the additional protocol changes triggered by the IRB review disagreements added any significant human subject protections.
Discussion: The study survey participants perceived significant variability in the IRB protocol review decisions, and they believed that the additional required edits triggered by IRB disagreement added few protections to the subjects involved. Consistent with the experiences of researchers in clinical and translational research, our study also found that the changes mandated by various IRBs may have caused time delays and increased study personnel time for revisions and resubmission without providing significant added value for investigators or subjects.
Significance: This research suggests an opportunity to reduce burdens and barriers to multi-site educational research by streamlining IRB review processes. The changes to the Common Rule and its guidance, including encouraging utilizing single IRBs for multi-site trials and delegating exempt decisions to investigators, may provide avenues for productive solutions.
Level of Audience: Mid-career
Focus of Presentation: Continuum
PRESENTER: Gerald Crites
AUTHORS/INSTITUTIONS: G. Crites, AU/UGA Medical Partnership, Athens, Georgia, UNITED STATES|R.D. Blanchard, University of Massachusetts Medical School – Baystate, Springfield, Massachusetts, UNITED STATES|A. Nagler, American College of Surgeons, Chicago, Illinois, UNITED STATES|L. Howley, Association of American Medical Colleges, Washington, DC, District of Columbia, UNITED STATES|
ABSTRACT BODY:

Purpose: Surveys are widely used in health professions education (HPE) research. A recent study of the three highest-impact journals in the field found that 52% of all original research studies employed at least one survey (1). Despite the widespread use of surveys, many HPE researchers are unfamiliar with the best practices of survey design (2). Survey items that are not carefully written can be difficult for respondents to answer, leading to measurement error. Ultimately, survey instruments containing such poorly written items may fail to provide researchers with reliable, interpretable data, thereby making validity arguments difficult to defend and results hard to interpret (3). What is unknown, however, is the extent to which HPE surveys violate established best practices in survey design. Our purpose was to begin filling this gap by assessing the quality of the items used in published HPE surveys.

Approach/Methods: We performed an analysis of all articles published in Academic Medicine, Medical Education, and Advances in Health Sciences Education in 2013. We selected articles that employed at least one self-administered survey, commonly called a questionnaire, and excluded surveys that were part of an interview or focus group. We then coded the items on each questionnaire using a scoring rubric developed from the literature (4,5). The rubric addressed four violations of established best practices in the design and visual layout of Likert-type survey items: multi-barreled language, agreement response options, unlabeled response options, and non-substantive response options (e.g., “N/A”) displayed together with substantive response options.

Results/Outcomes: Our search yielded 187 research articles that used at least one survey of any type; 36 articles (19%) met inclusion criteria and included a copy of the questionnaire for coding, yielding 37 unique questionnaires. Of these questionnaires, 65% had at least one multi-barreled item; 57% had at least one item using agreement response options instead of construct-specific options and, on average, 48% of the Likert-type items employed agreement response options; 42% had at least one item with unlabeled response options; and 19% had at least one item with non-substantive response options that were not set apart spatially from substantive response options. Altogether, 92% of the questionnaires contained at least one violation of best practices.

Discussion: Questionnaires are an important part of HPE research. These findings suggest that a substantial proportion of published HPE questionnaires violate established best practices in survey design. Although we do not know if or how such flawed items may have influenced respondents, we suspect, based on decades of empirical research, that these items could have been misinterpreted or otherwise resulted in less-precise data; such data can negatively impact the inferences researchers are able to make and threaten validity arguments (3).

Significance: To be effective research tools, survey instruments must be carefully designed (2). Results from this study, although limited by its small scale, suggest that published HPE questionnaires could benefit from more informed design. Ultimately, failure to follow best practices could negatively impact HPE investigations, the majority of which employ survey methodology. Potential solutions to the problem of poorly designed HPE questionnaires will be discussed in the presentation.


Level of Audience: Mid-career

Focus of Presentation: UME, GME, CME
**Purpose**: Reporting of P-values, confidence intervals (CIs), effect sizes, and adequacy of statistical power remains incompletely characterized in health professions education research (HPER). We sought to characterize current and past reporting of these features by manually coding a systematic, random sample of HPER studies published over 30 years, and by automated text mining of all HPER abstracts since 1970. We also compared HPER with biomedical research reports.

**Approach/Methods**: We searched PubMed, Embase, and CINAHL on May 7, 2016, for comparative research studies. For manual coding, we randomly sampled 250 HPER reports published in 1985, 1995, 2005, and 2015, and 100 biomedical research reports published in 1985 and 2015. Two reviewers abstracted information on participants, study designs, P-values, CIs, effect sizes, and power analyses. For automated text mining, we identified all HPER reports published 1970-2015, and used a computer algorithm to identify P-values and CIs in abstracts.

**Results/Outcomes**: In 2015, P-values were reported in 69/100 (69%) abstracts and 94 (94%) main texts. CIs were reported in 6 abstracts (6%) and 22 (22%) main texts. Most P-values were ≤.05 (87% in abstracts, 77% in main texts). 12 studies (12%) reported a prospective power analysis, and 9 met their target sample size. From 1985-2015, the proportion of abstracts reporting at least 1 P-value increased (odds ratio [OR] 2.00 per decade; 95% CI 1.62, 2.46) but the proportion of main texts did not (OR 0.87; 95% CI 0.59, 1.29). The proportion of abstracts reporting a CI did not change significantly over time after Bonferroni correction (OR 2.87; 95% CI 1.04, 7.88) whereas main texts reporting a CI increased (OR 1.96; 95% CI 1.39, 2.78). Only 60/165 two-group HPER studies (36%) had ≥80% power to detect a between-group difference of 0.5 standard deviations. The median Type I error (i.e., the probability that at least 1 P-value will be <.05 if the null hypothesis is true) was 0.54, yet only 8 studies (3%) adjusted the threshold of statistical significance. Comparison with biomedical research reports revealed similar reporting of P-values, but more frequent use of CIs in biomedicine. In automated text mining of 56,440 HPER abstracts, 14,867 (26.3%) reported a P-value, 3024 (5.4%) reported a CI, and reporting of P-values and CIs increased from 1970 to 2015.

**Discussion**: P-values are ubiquitous in HPER, CIs are rarely reported, and most studies are underpowered. Nearly all reported P-values would be considered "statistically significant" using the nominal threshold of .05. Type I error is high due to multiple hypothesis testing.

**Significance**: HPER investigators should plan hypothesis tests thoughtfully and in advance, only conduct tests they intend to report, avoid selectively reporting "significant" P-values, and report summaries of study data (estimates of effect size and variance) in addition to statistical test results. They should also estimate sample size requirements in advance, and use CIs in reporting results.


**Level of Audience**: Mid-career

**Focus of Presentation**: Continuum

**PRESENTER**: David Cook

**AUTHORS/INSTITUTIONS**: D.A. Cook, C.P. West, Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota, UNITED STATES|V. Pankratz, University of New Mexico Health Sciences Center, Albuquerque, New