The Inter-Rater Reliability of Technical Skills Assessment and Retention of Rater Training

Nada Gawad, MD, MAEd, [*,†,‡] Amanda Fowler, MD, FRCSC, [*,†,‡] Richard Mimeault, MD, FRCSC, [*,†] and Isabelle Raiche, MD, MAEd, FRCSC [*,†,‡]

[*]Division of General Surgery, Department of Surgery, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada; [†]The Ottawa Hospital, Ottawa, Ontario, Canada; and [‡]Department of Innovation in Medical Education (DIME), University of Ottawa, Ottawa, Ontario, Canada

BACKGROUND: The inter-rater reliability (IRR) of laparoscopic skills assessment is usually determined in the context of motivated raters from a single subspecialty practice group with significant experience using similar tools. The purpose of this study was to determine the IRR among attending surgeons of different experience and practices, the extent of rater training that is necessary to achieve good IRR, and if rater training is retained over periods of nonuse.

METHODS: In Part 1, 5 surgeons of different practice backgrounds assessed 3 laparoscopic cholecystectomy videos using the Global Operative Assessment of Laparoscopic Skills instrument. In Part 2, 2 of the surgeons assessed a total of 33 videos over 5 scoring sessions distributed across 6 months. They participated in 2 different training sessions, and retention was tested in the other 3 sessions. IRR was calculated for Parts 1 and 2 with an intraclass correlation (ICC) in a 2-way random-effects model.

RESULTS: The ICC for Part 1 was poor (ICC = 0.26). In Part 2, the ICC was highest after each training session (scoring #1 ICC = 0.76, scoring #3 ICC = 0.74). The ICC was not retained 1.5 months after the brief video-based training session (scoring #2 ICC = −0.17). The ICC was retained 2.5 months after the in-depth discussion training session (scoring #4 ICC = 0.70), but not 4.5 months later (scoring #5 ICC = 0.04).

CONCLUSIONS: Good IRR is not implicit among surgeons with varying backgrounds and experience. Good IRR can be achieved with different types of rater training, but the impact of rater training is lost in periods of nonuse. This suggests the need for further study of the IRR of technical skills assessment when performed by the wide variety of surgeon raters as is commonly encountered in the environment of postgraduate resident assessment. (J Surg Ed 76:1088−1093. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Laparoscopic skills assessment, Video-based assessment, Inter-rater reliability, Residents, Rater training, Simulation

COMPETENCIES: Medical Knowledge

INTRODUCTION

The introduction of competency-based medical education (CBME) emphasizes the need for valid and reliable assessment of trainees. In surgical training, one such area of need is the assessment of technical skills. While many assessment tools exist, the current literature tends to focus on determining validity and reliability in the development of the tool[1−3] where extraneous variables are minimized, and less so in clinical practice, where a variety of confounders exist.[4] One such confounder is the reliability between raters scoring trainees using these assessment tools.

At our institution, the laparoscopic skills of junior trainees are assessed using the Global Operative Assessment of Laparoscopic Skills (GOALS)[1] assessment tool on an ex vivo

Correspondence: Inquiries to Nada Gawad, MD, MAEd, Division of General Surgery, Department of Surgery, Faculty of Medicine, University of Ottawa, 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9, Canada; e-mail: ngawad@toh.ca

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

All authors drafted or critically revised the manuscript. Additional authorship contributions include:

- Study concept and design: Gawad, Fowler, Mimeault, Raiche;
- Acquisition of data: Gawad, Fowler, Mimeault, Raiche;
- Analysis and interpretation: Gawad, Raiche;
- Study supervision: Raiche.

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2019.01.001

porcine laparoscopic cholecystectomy model. To preserve trainee anonymity and improve feasibility of assessment, staff surgeons score resident videos remotely by watching a video recorded through the laparoscope. Asynchronous videobased rating has been demonstrated as an acceptable method of assessment for laparoscopic surgery.[4,5] Despite literature demonstrating generally excellent inter−rater reliability (IRR) with GOALS,[1,4] an informal review at our institution revealed large discrepancies between the video scores provided by different staff surgeons, raising concerns as to the utility of these discrepant scores for assessment within our training program. As such, we sought to investigate these observed discrepancies formally, and hypothesized suboptimal IRR across surgeon raters. We also hypothesized that any improvements made in IRR as a result of rater training would not be sustained over time.

A review of the literature on the IRR in the assessment of laparoscopic skills demonstrates that rater training tends to be brief,[1,4,6,7] and IRR tends to be good,[1,4,6−8] or not determined.[9,10] With respect to GOALS, much of the available literature originates from an institution where the tool is routinely used in research and/or in assessing trainees,[1,4,7,11] thus raters have pre-existing experience with the tool in addition to the minimal training provided.[4] In addition, raters are generally described as minimally invasive surgeons or fellows practicing in the same center of excellence,[1,4,7,11] which may predispose them to making similar judgments regarding acceptable variation in operative technique.[6]

Rater training is considered a necessary component in establishing the validity of an assessment tool, and is strongly recommended in the general education literature.[12] But the optimal method of training raters has yet to be defined.[4] Methods of rater training are often not described in detail, but when defined in technical skills literature, include familiarization with the assessment instrument items,[6,7] encouragement to use the full range of scores,[1,4] and instruction to consider factors such as level of difficulty, equipment, and lighting.[1,4] Other literature on IRR, however, describes more extensive rater-training techniques consisting of multiple days of theoretical training and practice scoring with feedback to raters from study investigators,[13] and hours of "norming sessions" involving prereading, group discussion around live ratings that have been preassigned an expert group consensus score, as well as one-on-one discussion with aberrant raters.[14]

Achieving good IRR is particularly integral with the introduction of CBME.[15] With respect to entrustable professional activities (EPAs) that must be documented by training programs, for feasibility reasons it is likely that 1 or 2 staff surgeons will assess each resident per EPA. Thus, the ratings provided by each staff must be uniform to provide each trainee with fair and accurate assessment throughout their training.

The purpose of this study was to determine the IRR between surgeons with different practice backgrounds and experience, and what extent of rater training is necessary to achieve good IRR between 2 academic surgeons with different experience rating videos of skill performance using the GOALS assessment tool. We also sought to determine if rater training is retained over periods of nonuse or if raters revert to their untrained rating habits.

METHODS

Procedure and Participants

All participants provided informed consent. Approval was obtained from our institutional ethics review board. Thirty-three procedures were videotaped of the dissection of an ex vivo porcine gallbladder from the liver bed as part of a laparoscopic cholecystectomy. The 33 procedures consisted of 10 PGY1, 5 PGY2, 5 PGY3, 5 PGY4, 5 PGY5, and 3 attending surgeon videos. Using GOALS, the videotaped procedures were assessed by attending surgeons. In Part 1, 5 surgeons with experience in laparoscopic cholecystectomy either through fellowship training in hepatobiliary surgery or minimally invasive surgery, and varying levels of experience using GOALS, rated 3 videos. Three videos were used for feasibility reasons, and the same 3 videos were used for all surgeons 5 surgeons. Following Part 1, in Part 2 we then had 2 of the surgeons (1 hepatobiliary and 1 minimally invasive) rate the remaining videos, for a total of 33 rated videos. One had ample experience using GOALS in the research context, and the other had no previous experience using GOALS beyond its use in our training program, thus mimicking the reality of surgeons involved in clinical training.

Study Design

We began by having all raters watch a 5-minute training video made by the study investigators, which provided an explanation of the scale and its components as well as video clips providing specific examples that would be considered consistent with scoring benchmarks (i.e., a "1," "3," and/or "5") for each respective item. After watching the training video, for Part 1, the 5 raters scored 3 videos independently. For Part 2, in addition to the training video, 2 raters met to briefly discuss their opinions of the examples shown in the training video in an attempt to achieve agreement on why the score was awarded. This served as our "minimal training" version analogous to previously published technical skills studies. Both raters subsequently proceeded to score videos independently (scoring #1), and their IRR was determined and the raters were informed of this result. Retention was then tested by having each rater score a

new set of videos 1.5 months later (scoring #2). IRR was again determined and the raters were informed of this result. One month later, the raters met again for an in-depth 3−hour discussion on their scoring opinions, biases, and questions with the goal of achieving standardization. To determine if this more extensive discussion had any impact on their scoring habits over the same period of time, they subsequently scored additional videos (scoring #3), IRR was determined, and raters were again informed. This process of scoring, calculating IRR, and informing the raters repeated 2 more times, first after another month (scoring #4), and finally after another 2 months (scoring #5). No further discussions about scoring were had between raters, and neither rater scored any other videos between scoring sessions. A timeline depiction of Part 2's rater training sessions, scoring sessions, and time between events is shown in Figure 1. Different videos were scored each time, and the number of videos scored at each session was determined by the number of videos available at the time based on logistics of participant scheduling and simulation center availability. The level of training of participant videos was distributed similarly across each scoring session, and each scoring session included videos spanning at least 4 levels of training to enable sufficient range of scores.

Assessment Tool

GOALS is a global rating scale consisting of 5 items: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy. Descriptive anchors guide the rater as to what is indicative of scores 1, 3, and 5. A modified version of GOALS was used, which omitted the autonomy item because residents performed the simulated procedure completely independently, with no attending surgeon present, and no assistance from the study investigator holding the laparoscopic camera. In addition, to preserve resident anonymity, there was no audio on the videotapes from which video raters could assess autonomy. Omission of the autonomy item has been previously studied with validity evidence supporting the use of GOALS for video-based laparoscopic cholecystectomy assessment.[16] As such, the total possible score was 20, which is the sum of the scores obtained from the 4 included items.

Videotapes

Videotapes were recorded directly from the laparoscopic camera with an anonymous study number, then digitized and transformed into Quicktime (version 10.4, Apple Inc.) videos using iMovie (version 10.0.9, Apple Inc.) for Macintosh. Each procedure was made into a separate video clip that started once the laparoscopic camera was inserted into the trainer box and was terminated once the gallbladder was free from the liver bed. Other than the start and stop times and omitted audio described, the videos were not edited. There were no views of the room or the resident operating. Raters were permitted to fast-forward or rewind the videotape using their expert judgment to decide what to view in detail.

Statistical Analysis

IRR was determined by calculating intraclass correlations (ICC) and their 95% confidence intervals using SPSS (version 24) based on a mean-rating (Part 1 $k = 5$, Part 2 $k = 2$), absolute-agreement, 2-way random-effects model. Because the purpose of GOALS assessment in our context and in the upcoming CBME era is both formative and summative assessment, cut-off points for the ICC were chosen as <0.5 = poor, 0.5 to 0.75 = moderate, 0.75 to 0.90 = good, >0.90 = excellent reliability.[17]

RESULTS

Part 1

Among the 5 surgeons, the mean score was 12.4 and the 95% confidence intervals were −2.31 to 0.98. The ICC was 0.26. This result prompted Part 2.

Part 2

The mean scores and ICC for each of the 5 scoring sessions are shown in Table 1 and Figure 2. The ICC was the highest in the scoring session that occurred after the training video + brief discussion session (scoring #1) and after the long in-depth discussion training session (scoring #3). The ICC after these training sessions was 0.76 and 0.74, respectively.

In terms of retention, the ICC was not preserved 1.5 months after the first training session (scoring #2), when it dropped to −0.17. However, retention was preserved 2.5

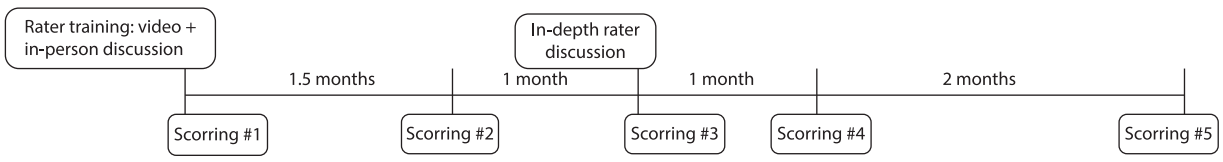

FIGURE 1. Timeline depiction of rater training and scoring sessions.

TABLE 1. Inter-Rater Reliabilities Between the 2 Raters for Each Scoring Session (Part 2)

Scoring Session	Number of Videos	Mean Score	ICC	95% CI
1	3	11.8	0.76	−0.35 to 0.99
2	5	11.2	−0.17	−2.70 to 0.85
3	5	9.6	0.74	−3.14 to 0.97
4	13	9.5	0.70	0.91 to 4.40
5	7	12.3	0.04	−0.88 to 0.75

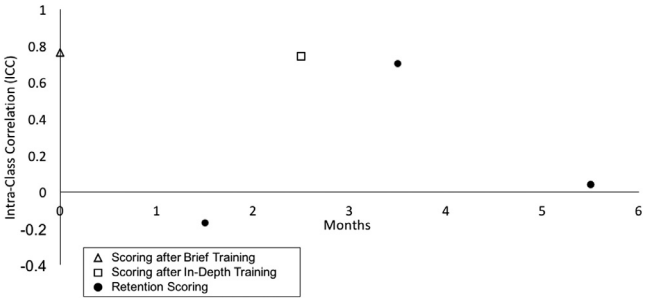

FIGURE 2. Inter-rater reliability over time (Part 2).

months after the in-depth rater discussion (scoring #4) at 0.70. The ICC was not preserved 4.5 months after the in-depth rater discussion (scoring #5) at 0.04. There were no trends observed across the scores of videos rated in scorings #4 or #5 to suggest rater drift within either session.

DISCUSSION

This study demonstrates that among surgeons with varying practice backgrounds and experience, minimal training does not ensure good IRR. In addition, both a training video with a short discussion as well as a long in-depth discussion were effective in achieving good IRR between 2 surgeon raters. However, after 1.5 months of nonuse, the long in-depth discussion training was retained, whereas the video and short discussion training was not. This suggests that not all rater training is equal, and there is some benefit to more detailed training. Furthermore, the loss of good IRR 4 months after the in-depth training suggests that raters require retraining if they are not rating videos consistently. We hypothesize that since rater training involved discussion, the raters were achieving some extent of consensus on variation in what they consider acceptable during training, but then reverted to their untrained habits after extended periods of nonuse due to the stronger influence of their personal experience.

Given the consequences of resident assessment data in an era moving toward CBME, it is insufficient to have well-developed assessment tools without employing equal effort in training the raters who use them.[4] While it is known that rater training improves IRR,[7] there is a paucity of literature on rater training in technical skills assessment. For this study, we chose the dissection of an ex vivo porcine gallbladder off a liver bed as it can be performed by residents of all levels thus allowing a wide range of scores, and because raters of any subspecialty who maintain a general surgery practice through on-call or acute care coverage could reasonably be considered experts. Yet these different practice backgrounds provide different lenses through which the surgeons rate technical skill, and without some standardization in what is considered "good" or "bad," the assessment data provided to a resident or their program are of little value. While there is no literature describing what factors may influence IRR in laparoscopic skills assessment, we noted that in both parts of this study, our surgeon raters consist of both genders, have a greatly different number of years in practice, practice at different hospitals within our institution, are engaged in different areas of research, and belong to different clinical subspecialties. On the other hand, all surgeon raters are considered experts technically and are highly regarded in our program as excellent clinical educators. The demonstrated dependency of IRR on sufficient and recent rater training represents a form of construct-irrelevant variance: score variance due to factors unrelated to the competency being assessed.[21] Construct-irrelevant variance because of inadequate rater rating threatens the practical use of assessment tools such as GOALS, an especially important consideration in the era of CBME. This study represents the reality of the wide variety of surgeon assessors within any residency training program, as opposed to the often-published examples of highly motivated raters within 1 practice group who have inherent familiarity with assessment tools.[4]

Three studies from the same research group describe rater training in technical skills assessment using GOALS use videotaped operating room and/or simulated laparoscopic inguinal hernia repair,[7] and live operating room[1] and videotaped[4] dissection of a gallbladder from the liver bed during laparoscopic cholecystectomy. All studies include rater training described as "brief."[1,4,7] Reported IRR (ICC) between 2 trained raters was 0.88 to 0.90,[7]

0.89,[1] and 0.39 to 0.94,[4] respectively. It is important to note that raters are reported to be minimally invasive fellows, attending surgeons, and/or researchers[1,4,7] who usually have extensive experience using GOALS. The single study with a wide range of IRRs points out that despite equal rater training, the bottom half of reported ICCs were by raters with no previous experience using GOALS.[4] Accordingly, a recent study performed by a different group with minimal previous experience using GOALS for laparoscopic cholecystectomy reported low IRR (ICC = 0.37).[16] As in our study, the lack of inherent familiarity with the GOALS tool suggests the need for extensive rater rating. The correlation of exposure to assessment method and reliability of ratings has similarly been demonstrated across other assessment tools for video-based laparoscopic surgery.[18] As demonstrated in our results, this highlights the need for the appropriate training of surgeon raters when used outside of the context of a research study with highly motivated and experienced raters[4] and offers an explanation as to why our IRR is so different from similar existing studies.

This study is not without its limitations, however. First, to highlight the good and poor IRR in the longitudinal part of our study (Part 2), we only used 2 raters. In ideal settings, the more raters scoring each video, the less the impact of 1 rater off-setting the reliability of the group. However, the implication of such a discordance of scores for a resident being assessed only twice on a given procedure is important, because it is more likely to represent the reality of practice. More observations than are feasible may be needed to achieve minimal levels of reliability for a given EPA.[19,20] In addition, Part 1 demonstrated similarly poor IRR even with 5 surgeons, suggesting that increased numbers of raters may in isolation still be insufficient to achieve good IRR. This further highlights the importance of rater training in practice environments where the addition of more raters may not be feasible.

Next, rater drift was not formally examined within scoring sessions. Although index videos were not included within each scoring session to purposefully assess for rater drift, this limitation was mitigated by raters taking breaks between video scoring. Furthermore, scoring trends were compared for videos rated early versus late in each of scoring sessions #4 (13 videos) and #5 (7 videos).

Finally, the longitudinal design of this study may have contributed to its limitations. In Part 2, it is possible that there was better retention of training after the in-depth training session because of improved familiarity with the tool after having scored 13 videos prior to scoring #4 as opposed to just 3 videos prior to scoring #2. We tried to mitigate this influence by increasing the length of time between the in-depth training and the retention scoring sessions. Future studies would ideally randomize raters into 2 groups receiving different training. Given the ultimate poor IRR in scoring #5; however, it is noteworthy that increasing familiarity with scoring videos is not enough to ensure reliable rating.

CONCLUSION

Our study is a preliminary step in demonstrating that good IRR in trainee assessment is not implicit among the wide variety of surgeon raters in practice, but can be achieved with sufficient rater training with variable retention. In practical terms, our results suggest that there is a need for training and retraining of attending surgeons who are not assessing resident videos for several months at a time to ensure reliable assessment. Further study of IRR in technical skills assessment should employ the wide variety of surgeons that realistically participate in resident training.

REFERENCES

1. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190:107–113. https://doi.org/10.1016/j.amjsurg.2005.04.004.

2. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa surgical competency operating room evaluation (O-SCORE). *Acad Med*. 2012;87:1401–1407. https://doi.org/10.1097/ACM.0b013e3182677805.

3. Martin J, Regehr G, Reznick RK, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278 http://journals2.scholarsportal.info.proxy.bib.uottawa.ca/pdf/00071323/v84i0002/273_osaotsfsr.xml.

4. Vassiliou MC, Feldman LS, Fraser SA, et al. Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. *Surg Innov*. 2007;14. https://doi.org/10.1177/1553350607308466.

5. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg*. 2008;247:372–379. https://doi.org/10.1097/SLA.0b013e318160b371.

6. Dath D, Regehr G, Birch D, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc*. 2004;18:1800–1804. https://doi.org/10.1007/s00464-003-8157-2.

7. Kurashima Y, Feldman LS, Al-Sabah S, Kaneva PA, Fried GM, Vassiliou MC. A tool for training and

evaluation of laparoscopic inguinal hernia repair: the Global Operative Assessment of Laparoscopic Skills-Groin Hernia (GOALS-GH). *Am J Surg*. 2011;201:54–61. https://doi.org/10.1016/j.amjsurg.2010.09.006.

8. Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. 2003;327:1032-1037. doi:10.1136/bmj.327.7422.1032.

9. Hogle NJ, Liu Y, Ogden RT, Fowler DL. Evaluation of surgical fellows' laparoscopic performance using Global Operative Assessment of Laparoscopic Skills (GOALS). *Surg Endosc*. 2014;28:1284–1290. https://doi.org/10.1007/s00464-013-3324-6.

10. Gumbs A, Hogle N, Fowler M. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007;204:308–313. https://doi.org/10.1016/J.JAMCOLLSURG.2006.11.010.

11. Vaillancourt M, Ghaderi I, Kaneva P, et al. GOALS-incisional hernia: a valid assessment of simulated laparoscopic incisional hernia repair. *Surg Innov*. 2011;18:48–54. https://doi.org/10.1177/1553350610389826.

12. National Research Council. How People Learn. Washington, DC: National Academies Press; 2000. https://doi.org/10.17226/9853.

13. Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the structured clinical interview for DSM-IV axis I disorders (SCID I) and axis II disorders (SCID II). *Clin Psychol Psychother*. 2011;18:75–79. https://doi.org/10.1002/cpp.

14. Cushing Weigle S. Using FACETS to model rater training effects. *Lang Test*. 1998;15:263–287. https://journals-scholarsportal-info.proxy.bib.uottawa.ca/pdf/02655322/v15i0002/263_uftmrte.xml. Accessed December 7, 2017.

15. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010;32:676–682. https://doi.org/10.3109/0142159X.2010.500704.

16. Kramp KH, van Det MJ, Hoff C, Lamme B, JGM Veeger N, Pierie J-PE. Validity and reliability of global operative assessment of laparoscopic skills (GOALS) in novice trainees performing a laparoscopic cholecystectomy. *J Surg Educ*. 2015;72:351–358. https://doi.org/10.1016/j.jsurg.2014.08.006.

17. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

18. Tadashi M, Hiroomi K, Akihiro K, et al. Reliability of laparoscopic skills assessment on video: 8-year results of the endoscopic surgical skill qualification system in Japan. *J Endourol*. 2014;28:1374–1378.

19. Meade LB, Borden SH, Mcardle P, Rosenblum MJ, Picchioni MS, Hinchey KT. From theory to actual practice: creation and application of milestones in an internal medicine residency program, 2004−2010. *Med Teach*. 2012;34:717–723. https://doi.org/10.3109/0142159X.2012.689441.

20. Sherbino J, Kulasegaram K, Worster A, Norman GR. The reliability of encounter cards to assess the CanMEDS roles. *Adv Health Sci Educ Theory Pract*. 2013;18:987–996. https://journals-scholarsportal-info.proxy.bib.uottawa.ca/pdf/13824996/v18i0005/987_troectatcr.xml. Accessed June 12, 2018.

21. Downing SM, Haladyna TM. Assessment in Health Professions Education. Downing SM, Yudkowsky R, eds. New York: Taylor & Francis; 2009. https://books.google.ca/books?id=7PyOAgAAQBAJ&pg=PA21&source=gbs_toc_r&cad=3#v=onepage&q=constructirrelevant.variance&f=false. Accessed November 2, 2018.