Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 8, June 2023 ISSN 1531-7714

Examination of the Aggregate Scoring Method in a Judgment Concordance Test

Marie-France Deschênes, *Université de Montréal, Québec*
Éric Dionne, *University of Ottawa, Ontario*
Michelle Dorion, *University of Ottawa, Ontario*
Julie Grondin, *University of Ottawa, Ontario*

The use of the aggregate scoring method for scoring concordance tests requires the weighting of test items to be derived from the performance of a group of experts who take the test under the same conditions as the examinees. However, the average score of experts constituting the reference panel remains a critical issue in the use of these tests. This study aims to examine the distribution of panelists' scores on the judgment concordance test (JCT) using the aggregate scoring method. A test composed of 32 items was developed and completed by 14 experts. The mean scores of the experts were calculated based on whether their choices of response categories for the 32 JCT items were included or excluded. Descriptive statistics were conducted. The mean scores of the experts showed a difference of 5.76%, depending on the approach used. The approach that excludes the experts' response category choices was found to be more penalizing (76.16%±8.9) than the method including their own choices (81.92%±8.1). It is recommended that researchers make their computational approaches explicit in addition to outlining the distribution of expert results retained for the purpose of determining scores in the concordance tests. Further research is required to refine our understanding of the quality of score-setting in this type of test.

Keywords: evaluation, concordance test, aggregate scoring, judgment, minority French-speaking communities

Introduction

The judgment concordance test (JCT) is an assessment tool increasingly used in initial and continuing health education programs. The JCT is a subcategory of script concordance test (SCT); while the JCT assesses professional judgment, the SCT assesses clinical reasoning. This implies the existence of differences in the format and content of the test items to prompt consideration of micro-decisions in uncertain or complex professional situations. However, both tests include the same design steps, cognitive tasks sought for each item, and processes for determining the candidates' scores. Regarding this process, the JCT and SCT both utilize the aggregate scoring method, in which the weighting of the test items is derived from the performance of a group of experts who complete the test under the same conditions as the examinees do (Norcini et al., 1990; Norman, 1985). There have been questions raised regarding the appropriate use of this method. In light of this, we studied the distribution of expert scores on the JCT using the aggregate scoring method before conducting the test with candidates.

The JCT and SCT are both composed of a series of vignettes that outline authentic situations commonly encountered in professional practice (Dory et al., 2012; Fournier et al., 2008; Lubarsky et al., 2011; Lubarsky et al., 2013). These situations are deliberately ill-defined,

and it is difficult to determine a response with a high degree of certainty. Each situation is followed by items related to an explanatory hypothesis (i.e., a plausible interpretation of the situation) or an intervention hypothesis (i.e., a probable intervention in this situation). Each hypothesis is then followed by a new piece of information. The cognitive task prompted by each item is to consider the effects of the new information on the suggested hypothesis. In other words, does the new information minimize, enhance, or have no effect on the hypothesis? Thus, Figure 1 presents an item in a concordance test in a generic form: if you think . . . and then you notice . . . the new information makes the hypothesis . . ., with each situation being followed by one or a series of items.

When assessing professional competencies such as professionalism and ethical judgment (Foucault et al., 2015), the item format of the JCT differs slightly. The cognitive task south is to judge behaviors and their consequences. The vignette situation is again briefly described, and the candidate is prompted to make a decision based on limited, complex, or even contradictory information. However, in this case, they must judge whether or not the described behavior is relevant on a Likert scale with four points, ranging from very relevant to irrelevant. The scale has no median value to compel examinees to make a judgment (Foucault et al., 2015).

The concordance tests are based on script theory. The term "script" refers to networks of knowledge structured and organized in practitioners' long-term memory (Charlin et al., 2000). These knowledge architectures allow experienced practitioners to identify and understand salient data or key elements of a problematic situation and generate hypotheses to resolve it. Practitioners then directly collect data to reinforce, minimize, or prioritize these hypotheses. Concordance tests are designed to mimic this cognitive hypothetico-deductive process of reasoning, involving several micro-judgments, that is, inferences and comparisons between the candidate's expectations of the situation presented and the different hypotheses suggested in the concordance test items (Charlin et al., 2000). These items are first answered by members of a reference panel composed of experts from the field. The notion of concordance in these tests implies that we can study candidates' micro-judgments by comparing them to answers previously provided by the panelists (Dory et al., 2012; Fournier et al., 2008; Lubarsky et al., 2011; Lubarsky et al., 2013).

The scoring process of a concordance test is complex and involves two steps (Dionne et al., 2017). First, panelists respond to the test items individually without consulting peers or reference books. The frequency distribution of the response categories as determined by the panelists is then used to determine the candidates' scores. The score calculated for each item reflects the degree of agreement between the candidates' and panelists' responses (Dory et al., 2012; Fournier et al., 2008; Lubarsky et al., 2011; Lubarsky et al., 2013), although the panelists may not necessarily agree on the choice of response categories.

Figure 1. Item format in a concordance test

Short case-based situation [...]		
If you think to …	And then, you notice …	The new information makes the hypothesis
[...]	[...]	☐ Extremely weakened; ☐ Weakened; ☐ Unchanged; ☐ Strengthened; ☐ Extremely strengthened
↑ *Hypothesis*	↑ *New information*	↑ *Micro-decision*

Table 1. Aggregate scoring method for establishing scores in a concordance test

	Response Category (n=5)				
	1	2	3	4	5
1. Identify the number of panelists in each response category to determine the modal response category (i.e., category 1 in this example)	7	1	1	1	0
2. Divide the number of panelists for each response category by the number of panelists who endorsed the modal response (i.e., divide by 7 in this example)	7/7	1/7	1/7	1/7	0/10
3. Determine the score of the candidates for each response category	1	0.14	0.14	0.14	0

Note: Response categories: 1: Extremely weakened; 2: Weakened; 3: Unchanged; 4: Strengthened; 5: Extremely strengthened

Considering the variability in the panelists' response choices, the test does not imply the presence of a right or wrong answer. However, patterns may sometimes emerge in panelists' responses (Dionne et al., 2017).

The use of the aggregate scoring method for scoring concordance tests was proposed by Charlin et al. (1998), and it has since been widely used in the scientific literature to determine concordance test scores (Dory et al., 2012; Fournier et al., 2008; Lubarsky et al., 2011; Lubarsky et al., 2013). Developed in the 1980s for other types of tests, this method requires the weighting of test items to be derived from the performance of a group of experts who take the test under the same conditions as the examinees (Norcini et al., 1990; Norman, 1985). For each response category, the weight is calculated as the number of experts who chose the category divided by the number of experts who endorsed the modal category (the most selected category). Specifically, candidates receive the maximum score (of 1) if they choose the modal category determined by experts, whereas they receive partial credit (between 0 and 1) if they make a response category choice that was chosen by at least one expert. Finally, they receive no points (0) if they make a choice that no expert has chosen. A candidate's score is the sum of the weights of the response categories chosen for all test items (Norcini et al., 1990; Norman, 1985). Table 1 illustrates the steps involved in determining scores using the aggregate scoring method, with an example of a panel composed of 10 experts.

Dory et al. (2012) and Lubarsky et al. (2011) conducted literature reviews of the construction and implementation of JCTs and SCTs. Overall, both of the tests have been shown to accurately assess professional judgment and clinical reasoning, respectively. Moreover, the authors highlighted the robustness of the tests, as evidenced by a statistically linear progression of the results between groups with different levels of expertise; experts and candidates with more experience performed better than novices and students. Although the authors noted that the aggregate scoring method was unusual, they considered it appropriate for measuring professional judgment or clinical reasoning using concordance tests. They pointed out that studies that examined other scoring methods did not show significant gains compared with the aggregate scoring method. The rationale of Dory et al. (2012) and Lubarsky et al. (2011) was based on the internal consistency of the tests reported in the reviewed studies and the significant differences in the candidates' scores based on their level of expertise. Similarly, Fournier et al. (2008) and Lubarsky et al. (2013) provide guidelines for writing SCTs and JCTs, and considered the aggregate scoring method to be the most appropriate method for determining scores using these tests. Acknowledging that the use of this method remains a source of debate, the authors suggest that psychometric research is required to study JCTs and SCTs in conjunction with other competence assessment tools.

Although frequently used in studies aiming to determine scores in concordance tests (Dory et al., 2012; Fournier et al., 2008; Lubarsky et al., 2011), the aggregate scoring method has been criticized by some authors (Bland et al., 2005; Lineberry et al., 2013; Wilson et al., 2014). For instance, Bland et al. (2005) questioned the robustness of the SCT for assessing clinical reasoning due to the absence of a single correct answer, and argued that the scoring method does not consider the accuracy of the examinee's choice of response category. For example, an item for which all panelists answered "Strengthened" on a five-response-category scale assigns the same score of 0 to candidates who answered either "Extremely strengthened" or "Extremely weakened." Thus, a candidate who agrees with the panelists on the direction of the response category choice will receive the same score as one who does not agree. In Table 1, a candidate scores 0.14 points if they respond that a hypothesis is strengthened or weakened, if the majority of panelists consider it to be extremely weakened. Some researchers (Bland et al., 2005; Exantus, 2020; Lemay et al., 2010; Wilson et al., 2014) tested alternative scoring methods in concordance tests, including a method leading to dichotomous scores, a method with a penalty according to the distance from the expert's modal choices, a combination with the method of aggregate scoring, and a method with a penalty according to distance from the expert's modal choices.

The performance level or average score of experts constituting the reference panel remains a critical issue in the use of these tests as well. In short, it is essential to question their own concordance scores compared with those of their expert colleagues using this method before assigning scores to candidates. Based on a literature review performed by Exantus (2020), it appears that expert scores frequently fluctuate around 76.6 ± 5.3% with the aggregate scoring method. However, the approach used for calculations, as well as the detailed distribution of panelists' scores, has not always been reported in these studies. Thus, the aggregate scoring method may produce an overestimated score if experts' choices on items are retained when determining performance scores. The lack of openly available software to easily calculate experts' scores in a concordance test also raises reasonable doubt regarding the calculations made using this method. Although there are software similar to Excel for calculating candidate scores, there is no published tool that is readily available for calculating expert scores.

To the best of our knowledge, no previous studies have reported the overestimated scores generated by the aggregate scoring method when calculating panelist scores in a concordance test. Thus, this study aims to examine the distribution of panelists' scores on the JCT using the aggregate scoring method.

Materials and Methods

Design

A descriptive quantitative approach (Mangold & Adler, 2019) was used to describe this phenomenon. This research design enables the description of a phenomenon using quantified data. This is often characterized by the use of data collected through surveys or tests. In our study, the unique variable examined was the panelists' scores, which were calculated under two conditions: using their own choice of answers for each item of the JCT, or not. The distribution of the scores under these two conditions was then compared using descriptive statistics, such as the mean, standard deviation (SD), confidence interval (CI), minimum and maximum values, and range. The aim was to describe the magnitude of the differences between the results from the two different conditions for the procedure when calculating the aggregate scores. In short, we examined the preliminary statistical data behind the design of the JCT, specifically the panelists' scores, that is used to assign scores to candidates for each item's response categories. This process was included in a research and development study aimed at designing a validated JCT for assessing active offer competency.

Active offer refers to the behaviors of healthcare and social service professionals related to services offered in both official languages of Canada (English and French) in their workplaces or internships (Casimiro et al., 2018). The purpose of an active offer is to ensure that the services offered by professionals, whether orally, in writing, or electronically, are easily accessible in both official languages in areas with a significant Francophone population. The quality of service provided in the official language of the client's choice is at the heart of the active offer concept. This is manifested by a bilingual greeting in person or on the phone, visual identification that professionals are

bilingual, and the publication of bilingual documents (brochures, websites, etc.). In the context of the research project, active offer targets the sensitivity and commitment of healthcare professionals to the delivery of healthcare services in minority French-speaking communities. Considering this, we chose to use the JCT because the test was intended to assess the behaviors of healthcare professionals with regard to demonstrating professional values and effective active offer in common professional practice situations.

Development and preliminary validation of the Judgment Concordance Test. The JCT was developed by research team members (n=6) and experts (n=14) in active offer from Canada using a research and development process (see Figure 2).

As a first step, an in-depth study of the competency framework was carried out so that all members could become aware of the essential resources available for healthcare workers in terms of active offer. These resources (knowledge, skills, attitudes, and values conveyed) are listed in the *Consortium national de formation en santé* (CNFS), a group of universities and colleges that offer French-language programs in various health disciplines. Next, the professional situations that health professionals encounter were recorded by three members of the research team and used to solicit several resources listed in the competency framework. These situations were presented and discussed through group discussions with experts in the domain (n=6) who collaborated in the development of the situations. These expert collaborators were partners from the CNFS. The goal was to ensure consistency between the content of the JCT and the solicited resources. Thirty-two professional active offer situations encountered by health professionals were composed and professionally revised for language. Figure 3 presents an example of

Figure 2. Stages of JCT development

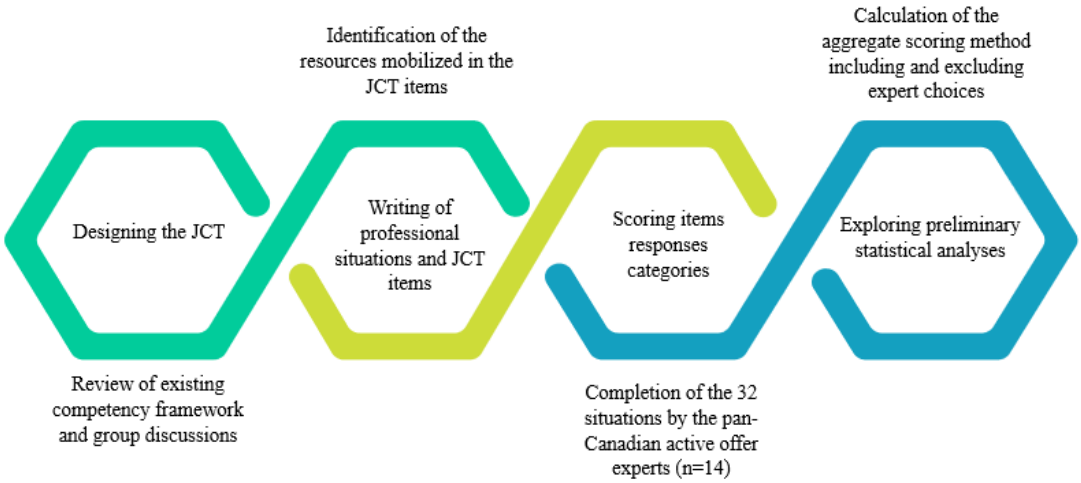


Figure 3. Example of an active offer JCT scenario

Remy is a physiotherapist in a private clinic in Moncton, New Brunswick. His patient, Mrs. Smith, is receiving physiotherapy after hip surgery. The meetings are conducted in French.	
Remy decides to...	**Following the intervention...**
… give the patient a copy of the homework program in French.	... Mrs. Smith asks him for the documentation in English. Her husband is an English speaker so he can help her do her exercises.
How would you rate Remy's intervention? ☐ : Very relevant; ☐ : Relevant; ☐ : Not very relevant; ☐ : Irrelevant	

an active offer situation, followed by an item for judging the relevance of the intervention. Each JCT item was coded in a blueprint representing the competency resources mobilized in the test situations. Once the JCT content was transcribed into a digital environment, the expert panelists completed the questions to measure their scores and thus determine the candidate item scores. Preliminary statistical analyses were then conducted based on the responses provided by the expert panelists.

The participants in this study were panelists, that is, peer-recognized experts in active offer. Solicited through network sampling, the participants were from various regions of Canada. The CNFS collaborators and the research team created a list of potential candidates. The experts were then contacted via email to participate in the study. Those who expressed interest in participating in the study were provided with a web link on the Moodle platform, where simple instructions were presented through a video, along with the 32 JCT professional situations. Participants were given three weeks to complete the JCT items. The panelists had not been previously involved in the drafting of JCT situations.

Data collection and analysis

The expert data were downloaded from the online platform for transcription using Excel spreadsheets. The data reported in this article refer to the quantitative data from the 14 experts, that is, their choices of response categories for the 32 JCT items. To facilitate data processing using the analysis software, the experts' response categories were recoded using numerical value, that is, interventions were deemed: 1 = irrelevant, 2 = not very relevant, 3 = relevant, and 4 = very relevant. From the compilation of the frequencies of the experts' response category choices for each item, the scores for each category were calculated using the aggregate scoring method. Then, with the scores determined for each response category, the expert scores were calculated, both including and not including their response category choices. An expert in educational measurement, co-author of this manuscript, was consulted for data processing and validation. An Excel spreadsheet was developed to calculate expert scores automatically. Descriptive statistics were then calculated (mean, SD, CI, minimum and maximum values, and range) using Excel spreadsheets and version 25 of the Statistical Package for the Social Sciences.

Following validation by the Research Ethics Board of the educational institute, an ethical review was not required in the context of an expert consultation process for the design of an educational tool. All experts participated freely in this project, and their data were anonymized for the analyses.

Methods

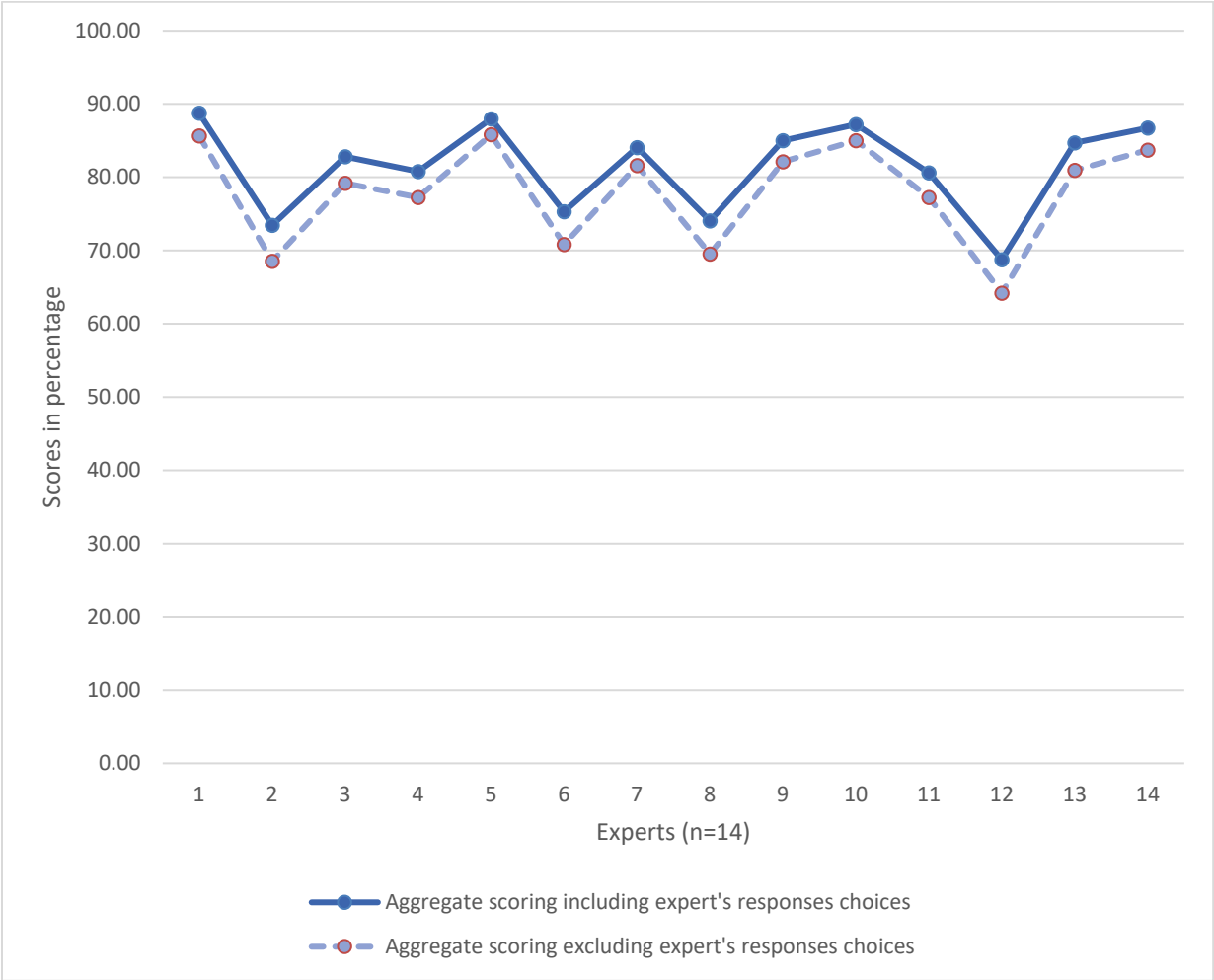
Fourteen experts participated in this study. Table 2 shows the mean scores of the experts based on whether their choices of response categories for the 32 JCT items were included or excluded. The mean scores of the experts showed a difference of 5.76%, depending on the approach used. The approach that excludes the experts' response category choices was found to be more penalizing (76.16%±8.9) than the method including their own choices (81.92%±8.1).

Figure 4 shows the different scores obtained by experts (n=14) using the aggregate scoring method,

Table 2. Average scores of experts when including or excluding their choices of response categories for JCT items

	Mean score of experts (n=14) including their choice of response categories	Mean score of experts (n=14) excluding their choice of response categories
Mean +/- SD	81.92 % ± 8.11	76.16 % ± 8.91
95% CI, lower bound	77.24	71.02
95% CI, upper limit	86.60	81.30
Maximum score	89.38	85.31
Minimum score	59.06	52.19
Range	30.32	33.12

Figure 4. Comparison of scores with the inclusion or exclusion of each expert's response choices



including and excluding their response category choices. The Appendix shows the detailed scores for each expert (n=14), that is, the raw JCT scores and percentage scores. The smallest percentage differences in scores were observed for Experts #1 and #5, that is, 4.07% between the two score calculations, whereas Expert # 2 had the largest difference (7.5 %).

Discussion

The level of performance or average scores of the experts constituting a reference panel in a concordance test remains a critical issue in the assessment of candidates' responses to test items. This prompted us to further explore the aggregate scoring method for developing the JCT scores. We calculated the average score of the experts by including and excluding their own choices of JCT response categories. The results show that the removal of each expert's response category choices using the aggregate scoring method avoids overestimation of their scores, which was significant in our study. The aggregate scoring method allows for variability based on the choices of experts in setting scores and tends to reflect differences and trends that may be found in professional practice (Norcini et al., 1990). In the context of concordance tests, one criticism of the aggregate scoring method is that it may give weight to response categories considered to be less than ideal or opposite to the majority of the experts' response categories (Bland et al., 2005; Exantus, 2020; Lineberry et al., 2013; Wilson et al., 2014). Various possibilities for addressing this have been investigated, such as optimizing the test by considering the optimal number of items and number of experts required (Gagnon et al., 2005; Gagnon et al., 2009; Gagnon et al., 2011), optimizing the number of answer choice categories to reduce candidates' tendencies to avoid extreme category choices (See et al., 2014; Wan et al., 2018), using alternative scoring

methods (Bland et al., 2005; Lineberry et al., 2013; Wilson et al., 2014), and determining the degree of agreement between the experts (Blais et al., 2012).

However, to the best of our knowledge, no study has shed light on the extent of score overestimation when calculating the aggregate scoring method using concordance tests. Our results show that experts' scores may be inflated if their choice of category is retained when determining their own scores. Considering that the candidates' JCT scores depend on the experts' choices of item response categories, it is necessary to report the distribution of expert scores to ensure that comparisons with candidate scores are meaningful (Boulouffe et al., 2014). In some published studies, researchers reported that individual expert scores were calculated by removing their own item response choices (Cooke et al., 2015; Deschênes et al., 2011; Ducos et al., 2015; Lambert et al., 2009; Latreille, 2012), whereas other researchers mention that the experts' scores were calculated manually (Bursztejn et al., 2011). However, the details remain underreported or unclear. Thus, we cannot conclude beyond any doubt that the potential overestimation of expert scores has not been avoided in other contexts. However, a rigorous approach to explicit presentation of calculation procedures and scores obtained is essential for the quality of the conclusions reported in the studies.

The effects on the candidates' scores based on the use of differing approaches to calculating aggregated scores are not trivial. In previous research, an overestimation of expert scores (by including their own answer choices in the scoring process) could promote significant differences between expert and candidate scores. In this regard, the statistically significant difference between expert and candidate scores has often been documented as evidence of the robustness of the data collected from JCTs and SCTs (Dory et al., 2012; Lubarsky et al., 2011). For evaluation purposes, candidate scores may be erroneously lower, whereas expert scores may be inflated. Removing experts' choices when calculating their own scores made the method of calculating aggregate scores consistent between the groups (experts and candidates). In addition, this approach does not unduly penalize candidates' scores when compared to those of experts.

An analysis of the experts' average scores being conducted prior to assigning scores to the candidates is essential, particularly with the aim of establishing a standard setting for candidates that is consistent with the performance of the experts. Considering the variability in experts' response choices, the aggregate scoring method in concordance tests does not imply the presence of a dichotomous response, that is, right or wrong answers to the items. As a result, the experts do not have a perfect average performance of 100% (Lambert et al., 2009). Exantus (2020) reported that the average expert performance in concordance tests was frequently approximately 75%–80% with the aggregate scoring method. Setting performance standards is an important issue in health education programs (Tavakol & Dennick, 2017) and cannot be done haphazardly in such contexts. Although efforts have been made to create a pass mark to establish the minimum level of performance in concordance tests (Charlin et al., 2010; Linn et al., 2013), other methods under different measurement theories must be employed to refine our understanding of scoring quality in this type of test.

However, this study was limited to one set of data. As a result, the context may influence the results, as may the number of experts or items included in the JCT. It is also uncertain whether the distribution of expert responses on a scale influences their performance levels. The results of this study cannot be compared with others, because there are few to no studies with this research objective in the existing literature.

Concordance tests are interesting tools whose originality lies, among other things, in the involvement of experts in the design of feedback for candidates who answer the test items. Feedback, whether in the form of a score or a written comment, is directly dependent on the panel of experts. However, the selection of experts for the reference panel remains unclear (Gawad et al., 2021). The panel's constitution, the level of difficulty and discrimination of the items, and the distribution and average scores of the experts are elements that need to be explored before establishing scores for the candidates. It is recommended that researchers make their computational approaches explicit in addition to outlining the distribution of expert results retained for the purpose of determining scores in the concordance tests. Further research is required to refine our understanding of the quality of score-setting in this type of test. Alternative scoring modalities in concordance tests can also shed new light on the ability of this instrument to assess active offer

competency, including a comparison of the JCT results with other competency assessment tools.

Conclusions

This study examined the distribution of JCT expert scores using the aggregate scoring method. The results indicate that the removal of each expert's response category choice avoids the overestimation of their scores, which was found to be significant in our study. Thus, it is recommended that in the future, researchers make their computational approaches explicit, in addition to outlining the distribution of expert results retained, for the purpose of improving the determination of scores in concordance tests. Further research is required to refine our understanding of the quality of the score setting for this type of test.

References

Blais, J.-G., Charlin, B., Grondin, J., Lambert, C., Loye, N., & Gagnon, R. (2012). Estimation du degré d'accord entre des experts lors du calibrage d'un test de concordance de script avec le modèle à facette de Rash. In G. Raîche, K. Paquette-Côté, & D. Magis (Eds.), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation* (Vol. 2, pp. 139-161). Presses de l'Université du Québec.

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine*, *80*(4), 395-399.

Boulouffe, C., Doucet, B., Muschart, X., Charlin, B., & Vanpee, D. (2014). Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation. *Emergency Medicine Journal*, *31*(4), 313-316. https://doi.org/10.1136/emermed-2012-201737

Bursztejn, A. C., Cuny, J. F., Adam, J. L., Sido, L., Schmutz, J. L., de Korwin, J. D., Latarche, C., Braun, M., Barbaud, A., & Venereology. (2011). Usefulness of the script concordance test in dermatology. *Journal of the European Academy of Dermatology*, *25*(12), 1471-1475. https://doi.org/10.1111/j.1468-3083.2011.04008.x

Casimiro, L., Savard, J., Sauvé-Schenk, K., & Atchessi, N. (2018). Questionnaire d'aiguillage pour dépister les besoins de formation quant à l'offre active de services en français. *Reflets. Revue d'intervention sociale et communautaire*, *24*(2), 182-211. https://doi.org/10.7202/1053868ar

Charlin, B., Brailovsky, C., Leduc, C., & Blouin, D. (1998). The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education*, *3*(1), 51-58. https://doi.org/10.1023/A:1009741430850

Charlin, B., Gagnon, R., Lubarsky, S., Lambert, C., Meterissian, S., Chalk, C., Goudreau, J., & van der Vleuten, C. (2010). Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teaching and Learning in Medicine*, *22*(3), 180-186. https://doi.org/10.1080/10401334.2010.488197

Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medecine*, *75*(2), 182-190. https://doi.org/10.1097/00001888-200002000-00020

Cooke, S., Lemay, J., Beran, T., Sandhu, A., & Amin, H. (2015). Development of a Method to Measure Clinical Reasoning in Pediatric Residents: The Pediatric Script Concordance Test. *Creative Education*, *7*(6), 814-823. https://doi.org/10.4236/ce.2016.76084

Deschênes, M.-F., Charlin, B., Gagnon, R., & Goudreau, J. (2011). Use of a Script Concordance Test to Assess Development of Clinical Reasoning in Nursing Students. *Journal of Nursing Education*, *50*(7), 381-387 https://doi.org/10.3928/01484834-20110331-03

Dionne, É., Grondin, J., & Latreille, M.-È. (2017). Exploration des scores à un test de concordance de script sous la loupe de la modélisation de Rash. In E. Dionne & I. Raîche (Eds.), *Mesure et évaluation en éducation médicale. Regards actuels et prospectifs* (pp. 77-110). Presse des Universités du Québec.

Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: insights from a systematic review. *Medical Education*, *46*(6), 552-563. https://doi.org/10.1111/j.1365-2923.2011.04211.x

Ducos, G., Lejus, C., Sztark, F., Nathan, N., Fourcade, O., Tack, I., Asehnoune, K., Kurrek, M., Charlin, B., & Minville, V. (2015). The Script Concordance Test in anesthesiology: Validation of a new tool for assessing clinical reasoning. *Anaesthesia Critical Care Pain Medicine*, *34*(1), 11-15. https://doi.org/10.1016/j.accpm.2014.11.001

Exantus, J. (2020). *Comparaison des propriétés métriques des scores obtenus avec un test de concordance de script au regard de trois méthodes de détermination des scores* [Master, Université d'Ottawa, Ottawa, Ontario, Canada].

Foucault, A., Dubé, S., Fernandez, N., Gagnon, R., & Charlin, B. (2015). Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): A pilot study. *Medical Teacher*, *37*(10), 955-960. https://doi.org/https://doi.org/10.3109/0142159X.2014.970986

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: guidelines for construction. *BMC Medical Informatics and Decision Making*, *8*(1), 8-18. https://doi.org/10.1186/1472-6947-8-18

Gagnon, R., Charlin, B., Coletti, M., Sauve, E., & van der Vleuten, C. (2005). Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Medical Education*, *39*(3), 284-291. https://doi.org/10.1111/j.1365-2929.2005.02092.x

Gagnon, R., Charlin, B., Lambert, C., Carriere, B., & Van der Vleuten, C. (2009). Script concordance testing: more cases or more questions? *Advances in Health Sciences Education*, *14*(3), 367. https://doi.org/10.1007/s10459-008-9120-8

Gagnon, R., Lubarsky, S., Lambert, C., & Charlin, B. (2011). Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? *Advances in Health Sciences Education*, *16*(5), 601-608. https://doi.org/10.1007/s10459-011-9279-2

Gawad, N., Wood, T. J., Malvea, A., Cowley, L., & Raiche, I. (2021). The Impact of Surgeon Experience on Script Concordance Test Scoring. *Journal of Surgical Research*, *265*, 265-271. https://doi.org/https://doi.org/10.1016/j.jss.2021.03.057

Lambert, C., Gagnon, R., Nguyen, D., & Charlin, B. (2009). The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. *Radiation Oncology*, *4*(1), 1-7. https://doi.org/10.1186/1748-717X-4-7

Latreille, M.-E. (2012). *Évaluation du raisonnement clinique d'étudiantes et d'infirmières dans le domaine de la pédiatrie, à l'aide d'un test de concordance de script* [Master, Université d'Ottawa]. Ottawa, Ontario, Canada.

Lemay, J.-F., Donnon, T., & Charlin, B. (2010). The reliability and validity of a paediatric script concordance test with medical students, paediatric residents and experienced paediatricians. *Canadian Medical Education Journal*, *1*(2), e89-e95. https://dev.journalhosting.ucalgary.ca/index.php/cmej/article/view/36531

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education*, *47*(12), 1175-1183. https://doi.org/10.1111/medu.12283

Linn, A. M., Tonkin, A., & Duggan, P. (2013). Standard setting of script concordance tests using an adapted Nedelsky approach. *Medical Teacher*, *35*(4), 314-319. https://doi.org/10.3109/0142159X.2012.746446

Lubarsky, S., Charlin, B., Cook, D., Chalk, C., & Van der Vleuten, C. (2011). Script concordance testing: a review of published validity evidence. *Medical Education*, *45*(4), 328-338. https://doi.org/10.1111/j.1365-2923.2010.03863.x

Lubarsky, S., Dory, V., Duggan, P., Gagnon, R., & Charlin, B. (2013). Script concordance testing: from theory to practice: AMEE guide no. 75. *Medical Teacher*, *35*(3), 184-193. https://doi.org/10.3109/0142159X.2013.760036

Mangold, K., & Adler, M. (2019). Designing Quantitative Research Studies. In D. Nestel, J. Hui, K. Kunkler, M. W. Scerbo, & A. W. Calhoun (Eds.), *Healthcare Simulation Research. A Practical Guide* (pp. 169-174). Springer.

Norcini, J. J., Shea, J. A., & Day, S. C. (1990). The use of aggregate scoring for a recertifying examination. *Evaluation & the Health Professions*, *13*(2), 241-251.

Norman, G. R. (1985). Objective measurement of clinical performance. *Medical Education*, *19*(1), 43-47. https://doi.org/10.1111/j.1365-2923.1985.tb01137.x

See, K. C., Tan, K. L., & Lim, T. K. (2014). The script concordance test for clinical reasoning: re-examining its utility and potential weakness. *Medical Education*, *48*(11), 1069-1077.

Tavakol, M., & Dennick, R. (2017). The foundations of measurement and assessment in medical education. *Medical Teacher*, *39*(10), 1010-1015. https://doi.org/10.1080/0142159x.2017.1359521

Wan, M. S., Tor, E., & Hudson, J. N. (2018). Improving the validity of script concordance testing by optimising and balancing items. *Medical Education*, *52*(3), 336-346. https://doi.org/10.1111/medu.13495

Wilson, A. B., Pike, G. R., & Humbert, A. J. (2014). Analyzing script concordance test scoring methods and items by difficulty and type. *Teaching and Learning in Medicine*, *26*(2), 135-145.

Citation:
Deschênes, M. F., Dionne, E., Dorion, M., & Grondin, J. (2023). Examination of the aggregate scoring method in a judgment concordance test. *Practical Assessment, Research, & Evaluation*, 28(8). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/8/

Corresponding Author:
Marie-France Deschênes
Université de Montréal, Québec
Email: marie-france.deschenes@umontreal.ca

Appendix A.

Expert Scores and Descriptive Analyses Using the Two Approaches to Calculating the Combined JCT scores

	Initial scores (/32) with experts' choice	**Score %**	**Initial scores (/32) without the choice of experts**	**Score %**
Expert 1	28.60	89.38	27.30	85.31
Expert 2	23.40	73.13	21.00	65.63
Expert 3	27.00	84.38	24.90	77.81
Expert 4	27.20	85.00	25.50	79.69
Expert 5	28.20	88.13	26.90	84.06
Expert 6	25.80	80.63	24.00	75.00
Expert 7	28.20	88.13	26.70	83.44
Expert 8	25.50	79.69	23.50	73.44
Expert 9	25.40	79.38	23.20	72.50
Expert 10	28.40	88.75	26.90	84.06
Expert 11	25.30	79.06	23.30	72.81
Expert 12	18.90	59.06	16.70	52.19
Expert 13	27.10	84.69	25.20	78.75
Expert 14	28.00	87.50	26.10	81.56
	Mean	81.92	Mean	76.16
	SD	8.11	SD	8.91
	MAX	89.38	MAX	85.31
	MIN	59.06	MIN	52.19