
Article:
Leppink, Jimmie orcid.org/0000-0002-8713-1374 (2017) Evaluating the strength of evidence in research and education:The theory of anchored narratives. Journal of Taibah University Medical Sciences. pp. 284-290. ISSN 1658-3612

https://doi.org/10.1016/j.jtumed.2017.01.002

Journal of Taibah University Medical Sciences (2017) 12(4), 284–290

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com

Educational Article

Evaluating the strength of evidence in research and education: The theory of anchored narratives

Jimmie Leppink, Ph.D.

Department of Educational Development and Research, School of Health Professions Education, Maastricht University, Maastricht, The Netherlands

Received 15 November 2016; revised 12 January 2017; accepted 15 January 2017; Available online 24 February 2017

الملخص

إن البحوث التجريبية المبنية على مجموعة من المشاركين وتقييم كفاءة كل طالب، أو متدرب، أو مهني في موضوع محدد، تشترك في أمر واحد على الأقل؛ ينبغي تحديد الأدلة المؤيدة أو المضادة للفرضية بأخذ أجزاء مختلفة من الأدلة في الاعتبار بعناية وتكاملها لخلق قصة متماسكة بلا تناقضات، ولا أطراف غير مترابطة، ولا عناصر مفقودة. ولتأمين إطار تفكيري متماسك لهذه العملية، تقدم هذا المقالة نسخة معدلة من نظرية استُخدمت كنموذج لاتخاذ القرارات القانونية في الحالات الجنائية؛ نظرية الروايات الراسية. في هذه النظرية يقوم القضاة في قضية ما، بالحكم على جودة أجزاء من الأدلة، وما إذا كان من الممكن تثبيت تلك الأجزاء كروايات تشكل سلسلة من الأدلة تُمكن من التوصل إلى قرار لا يدع مجالا للشك بجُرم المتهم. توفر هذه المقالة أمثلة من المجال الطبي لإيضاح كيف يمكن لنسخة معدلة من هذه النظرية أن توفر إطارا فكريا للباحثين والمربين يكون فيها تقييم كل من البحث التجريبي والكفاءة في النهاية دائما حكما مهنيا نوعيا مبنيا على دمج مجموعة مختلفة من المعلومات النوعية والكمية.

الكلمات المفتاحية: الطب؛ التعليم الطبي؛ نظرية الروايات الراسية؛ قصة؛ دليل

Abstract

Empirical research based on groups of participants and assessment of the competence of individual students, trainees, and professionals in a given context have at least one thing in common: evidence in favour or against a hypothesis should be established by carefully considering and integrating various pieces of evidence to create a coherent story that has no contradictions, loose ends or missing elements. To provide a coherent framework for this process, this article introduces a modified version of a theory that has been used as a model of legal decision making in criminal cases: the theory of anchored narratives. In this theory, judges in a case judge the quality of pieces of evidence and whether these pieces of evidence can be anchored as narratives to form a chain of evidence that enables a decision beyond reasonable doubt regarding a suspect's guilt. This article provides examples from the domain of medicine to elaborate how a modified version of this theory can provide researchers and educators with a framework in which the assessment of both empirical research and competence is a qualitative professional judgement based on an integration of various sources of qualitative and quantitative information.

Keywords: Evidence; Medicine; Medical education; Story; Theory of anchored narratives

Introduction

Empirical research never occurs in vacuum; theory, potentially relevant previous research, setting, and interests and expectations emerging from ongoing intellectual dialogue and multilogue place a given empirical study in a particular context that has implications for the meaning of findings within and beyond the study. Likewise, assessment of the competence of students, trainees, and professionals occurs in a given context in which a variety of assessments

Corresponding address: Department of Educational Development and Research, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands.
E-mail: jimmie.leppink@maastrichtuniversity.nl

Peer review under responsibility of Taibah University.

Production and hosting by Elsevier

 http://dx.doi.org/10.1016/j.jtumed.2017.01.002

are undertaken. In fact, whether we consider research on clinical reasoning,[1] legal decision making in a criminal case,[2] research in education,[3] or the assessment of the medical competence of individual students, residents or professionals,[4] the Latin adagio *unus testis nullus testis* is key: decisions ought not be based on a single source of information. Hence, whether we consider the evidence derived from empirical research with groups of participants or the assessment of learning or performance (e.g., current competence or an increase in competence in a given time interval), evidence in favour or against any given hypothesis ought to be established through a careful consideration and integration of a variety of pieces of evidence into a coherent story that has no contradictions, loose ends or missing elements.

For instance, to make appropriate decisions with regard to the medical conditions and needs of a patient who reports acute and severe chest pain, clinicians and other individuals (e.g., nurses and residents) have to ask the right questions, perform physical examinations, and think about possible diagnoses and other steps while continuously monitoring a patient's blood pressure, pulse rate, and respiration.[5] In a research field, the meaning and implications of findings from a scientific study are established through the context in which a study has taken place, previous research relevant to the study at hand, and contemporary theory.[3] In the context of the assessment of medical competence, professionals have to arrive at well-founded decisions about an individual's competence in a given context using qualitative and quantitative indicators of learning and performance from a variety of sources, which include objective, structured clinical examinations (OSCEs),[1] progress test scores,[6] course exams, interviews, and feedback from supervisors, patients or others.

In each of the aforementioned contexts – clinical reasoning, criminal law practice, educational research or the assessment of medical competence – multiple pieces of evidence have to be integrated into a *chain of evidence* that provides a coherent study with which professional judges – clinicians, judges or jury members, researchers and educational practitioners, and medical assessors – can make well-founded decisions. These decisions can pertain to the health status and needs of a patient (i.e., clinical case), the guilt or innocence of a suspect (i.e., criminal case), implications of research findings for theory, future research and educational practice (i.e., educational research), and the competence of a student, resident or professional (i.e., assessment of medical competence), respectively. Whatever practice we consider—and whether we address mainly qualitative or predominantly quantitative information—nothing operates in a vacuum: context is key.

This article introduces a modified version of a theory that was developed by legal psychologists Wagenaar, Van Koppen, and Crombag[2] as a model of legal decision making in a criminal case: the theory of anchored narratives (henceforth: TAN). In TAN, judges subsequently judge the quality of pieces of evidence (i.e., stories) and whether these pieces of evidence can be anchored as narratives to form a chain of evidence (i.e., a coherent story) that enables judges to decide beyond reasonable doubt about a suspect's guilt or innocence. After a concise presentation of TAN, using a criminal case example, this article introduces a modified version of TAN for the context of educational research and assessment, and discusses this modified version in light of the contemporary validity frameworks of Kane[7] and Messick[8] as well as current views on workplace learning and assessment.[9–11] Given its resonance with these frameworks, TAN provides a framework for the evaluation of the strength of evidence – in favour or against a given hypothesis – which underlines that the assessment of both empirical research and competence is in the end always a qualitative professional judgement based on an integration of a variety of qualitative and quantitative information.

In TAN, a judge comes to a decision concerning the guilt or innocence of a suspect in two stages. At the first stage, individual pieces of evidence, handed over to the prosecution and defence, are judged in terms of plausibility and quality. Subsequently, at the second stage, these pieces of evidence are evaluated in terms of how well they can be integrated or anchored into facts, common sense, and related to other pieces of evidence at hand. Each individual piece of evidence can reach the second stage (i.e., that of anchoring) only if a good and plausible story can be provided with it, and successful anchoring requires that this story be integrated into a chain of evidence. Take the following example, adopted and modified from[3]:

> *Dr. X. is found dead – with a single shot through the forehead – in the backyard of his house, and forensic examination reveals a match in DNA between suspect Dr. Y. and a piece of cigarette found in the backyard of Dr. X.*

In essence, a DNA match provides a potentially decisive piece of evidence. That is, if we can provide at least one piece of evidence—and preferably several other pieces of evidence that provide a chain of evidence—pointing at Dr. Y. being at the crime scene at or around the time of the death of Dr. X., the DNA match can be considered sufficient evidence to put Dr. Y. in jail. This evidence can come from eyewitnesses who report that they observed Dr. Y. at the crime scene around the time of the critical event, from others who report that Dr. Y. was not at home or in office as expected around that time, from global positioning system (GPS) data from a mobile device, and even other sources. What these sources of evidence have in common is that they allow the DNA match to be anchored in a coherent story line that can support interpretations of the DNA match in terms of Dr. Y. killing Dr. X. At the same time, one solid piece of evidence that points either against the presence of Dr. Y. at the crime scene around the time of the critical event or against the involvement of Dr. Y in the death of Dr. X. in some other way may have the potential to take stories about the guilt of Dr. Y. off the table. Other pieces of evidence may need to be examined in order to discard, beyond reasonable doubt, alternative scenarios such as the piece of cigarette being collected and put at the crime scene by someone who wants to set up Dr. Y. or the piece of cigarette having been dropped by Dr. Y. at a previous meeting between Dr. X. and Dr. Y. in Dr. X.'s backyard. Figure 1 provides an example of TAN in this example case.

TAN not only provides a model to explain decision making in successfully solved cases but also provides a model to explain miscarriages of justice, as a miscarriage of justice can usually be explained in terms of a judge (or jury, for that matter) either misjudging the quality or plausibility of

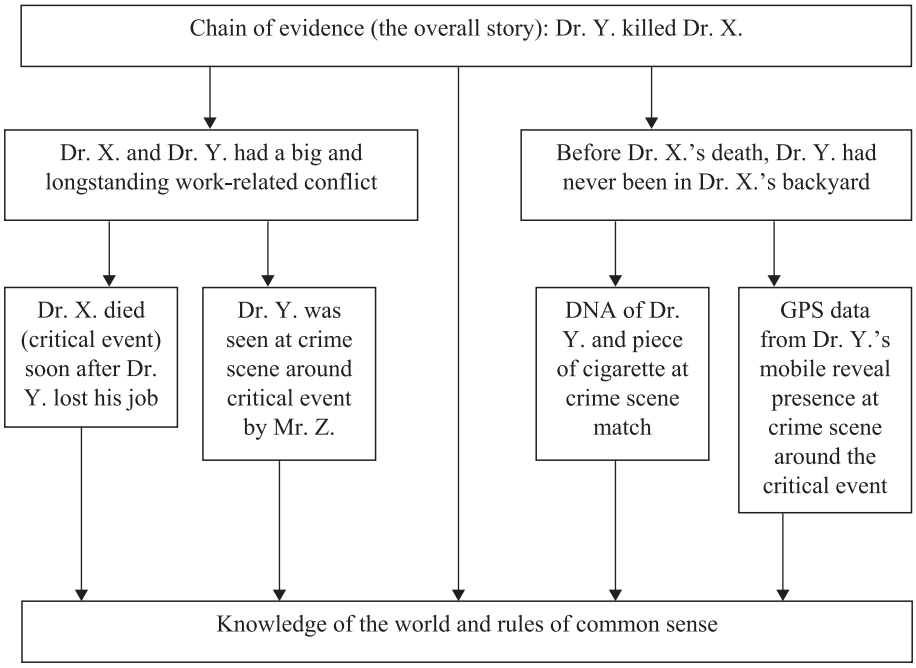

Figure 1: Concise depiction of the theory of anchored narratives through an example.

evidence (i.e., first stage) or inappropriately anchoring one or more pieces of evidence (i.e., second stage).[2]

Conceptual and measurement questions in educational research

In a modified form, TAN also applies to the evaluation of the strength of evidence for or against assumptions and hypotheses in empirical research. For example, take the widespread use of questionnaires, tests, and other psychometric instruments in the field of medical education. Whether we address knowledge progress testing,[6] the evaluation of clinical teaching,[12] seminar learning experiences,[13] culture in medical schools,[14] peer feedback,[15] cognitive load[5] or another construct, researchers frequently use psychometric instruments when they attempt to measure the construct(s) of interest. However, researchers at times forget that 'validating' an instrument is about a structured sequence of steps that together provide a chain of evidence[7,8] regarding the validity of an instrument (i.e., whether it measures what it is supposed to measure).

Logical relations between measures

The recent *Standards and Guidelines for Validation practices*[16] are heavily influenced by Messick's[8] unitary view of validity, in which construct validity or "the extent to which the relations among items, domains, and concepts support a priori hypotheses about the logical relations that should exist with other measures" (Ref. [16], p. 14) is the central component in validation research. Sources of evidence comprised in this component are test content, response processes, internal structure, relations to other variables and consequences of testing. Let us, to consider an example, focus on one of these — relations to other variables. This step is frequently omitted or not carried out rigorously: a careful study of how scores on an instrument under consideration relate to variables that are either known to measure the construct of interest or that — based on solid theory and previous research — can be expected to be strongly related to that construct.

Towards a stable instrument score (Y)

To develop an instrument, researchers need to carefully undertake a number of steps[17,18]: literature review, interviews and/or focus groups, synthesis of outcomes of the previous, item development, expert validation, cognitive pretesting and pilot testing. The latter step, *pilot testing*, should first aim at obtaining stable factors, that is, sets of items that can be grouped together as measuring the same underlying variables or *constructs* (e.g.,[5,12–15]). For this purpose, we can use factors analysis[19] or item response theory models[20] when we assume our constructs to be continuous (e.g., cognitive load), and we can use latent class analysis[21] or latent profile analysis[22] for constructs that we assume to be categorical (e.g., opinions about particular topics). Having stable factors is a necessary but not sufficient condition for being able to measure the constructs we are interested in. In other words, once we have stable factors, it is time to make the next move. For ease, this article focuses not on a multi-factor instrument[5,12–15] but on a single-factor instrument: in a hypothetical study, a group of researchers has developed (cf. [17,18]) five self-rating items that are supposed to measure cognitive load experienced by students while performing an OSCE, and factor analysis on different samples of students indicates the same

one-factor solution across samples. Hence, the ratings of the five items can be grouped together to obtain a single score Y for each respondent.

How Y relates to other variables (Z)

To examine whether this score Y really is an indicator of cognitive load experienced by students while performing an OSCE, we need to study carefully how Y (a) relates to variables that are known to measure the construct of interest or (*b*) relates to variables that — based on solid theory and previous research — can be expected to be strongly related to that construct, henceforth: Z. In the context of cognitive load, an example of (*a*) is found in *secondary task measures*[5]: students then perform a simple secondary task (e.g., pushing a button when there is a signal on a screen) that is unrelated to the primary task while they are carrying out the primary task. In the OSCE example, for instance, students could be instructed to focus on the OSCE and to push a button on a device within hand's reach every time the screen of that device lights up. Reduced accuracy and slower performance on the secondary task are known indicators of increased cognitive load.[5] At the same time, as an increased cognitive load from the primary task tends to result in more errors and/or slower performance on that primary task, an example of (*b*) is the performance on the primary task, in our case the OSCE.

Conversely, had the researchers' interest not been in a cognitive load instrument but in the development of a test for clinical reasoning during OSCE performance instead, with regard to (*a*) they could have done a literature search on known measures of clinical reasoning in OSCEs or similar examinations, and a cognitive load measure might have been included for (*b*).[5]

Manipulating an experimental treatment factor (X) to which Y and Z commonly respond

Observing a correlation between Y and Z (*a* and/or *b*) in itself does not provide many support regarding the assumption that Y is about our construct of interest; several unmeasured variables may resonate in the correlation observed. However, cognitive load theory has resulted in a wide variety of well-designed randomized controlled experiments that clearly indicate that, across settings, cognitive load and performance (i.e., Y and Z) commonly respond to manipulations in task complexity[5]: an increase in complexity typically results in an increased cognitive load and a decrease in performance (i.e., reduced accuracy and/or slower performance). In other words, if we design a randomized controlled experiment for the sole purpose of validation, in which we randomly assign a large group of students to either of two OSCE conditions that provide exactly the same instruction to students and vary *only* in the complexity of the case (i.e., *low* vs. *high* complexity), we should find the highest Y and lowest Z scores in the high-complexity condition. With this finding, we gain some support for the assumption that Y is about cognitive load; if the pattern of condition differences is a different one, we fail to gain that support. Figure 2 summarizes the story.

In Figure 2, Z is a known measure of cognitive load, X is an experimental manipulation which in cognitive load theory research has resulted in differences in cognitive load before,[5] and Y is the score of a new instrument that is intended to

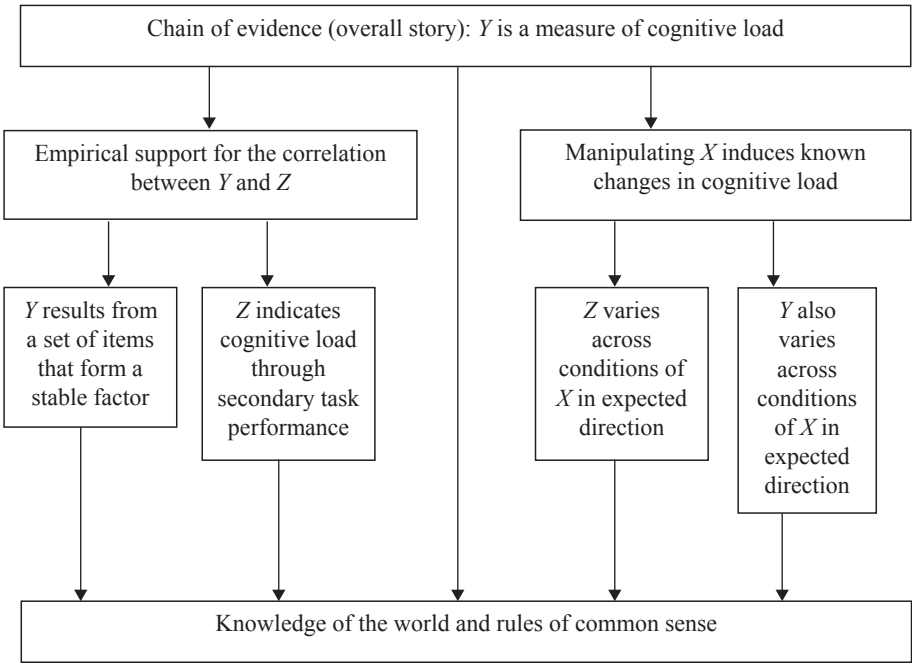

Figure 2: Anchoring narratives around the development of a measurement instrument: instrument score Y (i.e., assumed to measure cognitive load), known measure of the construct of interest Z (i.e., secondary task performance as a known measure of cognitive load), and a comparative variable X (in our case: the conditions in a randomized controlled experiment).

measure cognitive load. Given Z and X, the story that Y measures cognitive load can be anchored if Y correlates with Z as expected and covaries with Z across conditions of X as expected in cognitive load theory.

Relating Y to X and Z when experiments are not feasible

Of course, depending on what is being measured and in what context research occurs, randomized controlled experiments may ethically or logistically be hard to realize (e.g., [6,12–15]), meaning that X must be operationalized through careful comparisons of meaningful pre-existing groups and/or repeated measurements on Y and Z. Although a common association in this context has been that of expert-novice comparisons, Cook[23] reasons: "The major flaw is the problem of confounding: there are multiple plausible explanations for any observed between-group differences. The absence of hypothesized differences would suggest a serious flaw in the validity argument, but the confirmation of such differences adds little. As such accurate known-groups discrimination may be necessary, but will never be sufficient, to support the validity of scores" (p. 829). In well-designed randomized controlled experiments, confounding is rarely an issue; hence, findings in the hypothesized direction typically provide a much stronger case for X-to-Y causal inference when obtained in well-designed randomized controlled experiments than when coming from known-groups comparisons.

Nevertheless, there are cases when known-groups comparisons can be meaningful. One situation in which meaningful pre-existing groups are found easily is in the context of knowledge progress testing. In several medical curricula across the world, the progress test is administered three to four times every academic year. For instance, medical students in the Netherlands complete four progress tests in each of six academic years, yielding twenty-four progress test measurements for every student. As such, the progress test provides a powerful longitudinal tool for measuring knowledge progress throughout the medical curriculum, and this approach is based on a very solid theoretical framework.[6] Suppose, researchers have developed a new computerized adaptive progress test which is expected to provide about the same information about students' progress but with fewer test items. They randomly sample 600 medical students from the Netherlands – 100 students from each of six academic years – and have them perform the new test at a single point in time. Given that the existing progress test is a very well-established instrument, if the new test really measures knowledge progress among medical students, score Y derived from the test should highly correlate with score Z from the last regular progress tests that they completed in the context of their curriculum and should show differences between years X of a magnitude that is similar to that of scores Z on previous progress tests in the curriculum.

Unfortunately, some research contexts do not lend themselves to meaningful comparisons of pre-existing groups.[23] However, even then, we are not left with empty hands. In the OSCE example, for instance, if we have a series of cases that have to be completed by all students in a course, one might want to opt for counterbalancing the order of cases and randomly assign students to different orders. With three cases (A = easiest, B = more complex than A, C = more complex than B), for instance, that would yield six orders (ABC, ACB, BAC, BCA, CAB, CBA). Although this design is generally somewhat weaker than a randomized controlled experiment because all students undergoing all cases or conditions can come with the risk of carryover effects (i.e., one's performance on or mere confrontation with a current case affects one's performance on a next case), this design does enable collecting performance (Z) and cognitive load (Y) measurements after each case, and the counterbalancing enables accounting for order effects at least to some extent.[24]

Toward a chain of evidence: X–Y–Z interrelations confirmed, replicated, and meta-analysed

Generally speaking, when feasible, well-designed randomized controlled experiments that are designed for the sole purpose of validation can provide a stronger case for X-to-Y causal inference than studies that involve meaningful comparisons of pre-existing groups and/or repeated measures (e.g., counterbalanced order of conditions), given that 'validating' an instrument involves a structured sequence of steps[17,18] that together provide a chain of evidence.[7,8] Finally, whether we take Messick's unitary view of validity,[8] Kane's argument-based approach[7] or we apply TAN to this context, whichever of the aforementioned research designs we choose given our situation, a chain of evidence is not established in a single study involving X–Y–Z interrelations; we need replication studies[3,25] to enable the use of more powerful tools such as meta-analysis and systematic review. This also holds for studies the focus of which is not the development of instruments but for example which instructional formats work for which students.[3]

Developing an assessment story

In the previous section, we have seen that in the case of empirical research the evaluation of evidence pertains to integrating and anchoring a series of empirical studies with theory and context. Analogously, in the case of assessing competence, it is series of assessments that focus on a competence of interest that need to be integrated and anchored. For instance, assessing medical competence may include OSCE performance, progress test scores, course exams, interviews, and feedback from supervisors, patients or others during residency. Most of these are administered more than once, which is a good thing given that learning is the product of strings of experiences, combinations of non-linear trajectories, and social and cultural reifications[11] and given that every assessment has its limitations.[9,10] Like with empirical research using groups of participants (e.g., the development of a psychometric instrument), where both numbers and qualitative input from interviews and focus groups have a use (e.g., part of the support for the statement that "Y results from a set of items that form a stable factor" in Figure 2 may come from interview or focus group data), Figure 3 illustrates that assessment is a context where a considerable portion of information does not come from numbers but from verbal and, in certain cases, non-verbal data.

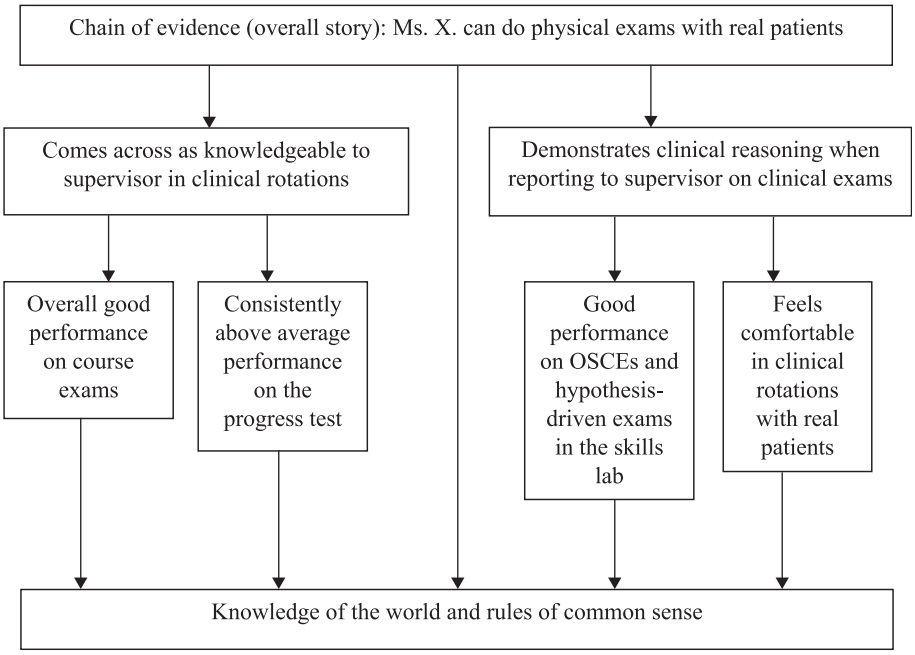

Figure 3: Developing a story about an individual's competence in a given context.

Unfortunately, some of the discourse around assessment has been led by some kind of quantitative-qualitative dichotomy, where a quantitative approach is equated with 'psychometrics' and a qualitative approach is equated with 'interpretivist'. Due to this phenomenon, psychometrics is associated with algorithms, reliability, and validity while an interpretivist (i.e., qualitative) approach to assessment is associated with triangulation of information, saturation of information, trustworthiness, and credibility. Such a dichotomous view on assessment ignores that psychometrics is one but not the only quantitative approach to assessment, fails to appreciate that some of the core assumptions underlying quantitative data in educational research and assessment are largely qualitative (e.g., the degree to which an assessment covers a domain of interest, the extent to which an assessment has face validity in the eye of an assessor or the one who is assessed), and disregards that a qualitative approach does not necessarily rule out the use of quantitative information. Statistics and psychometrics are not only about algorithms but also and perhaps largely about heuristics and a qualitative judgement at all stages. At the same time, a qualitative approach to assessment does not exclude numerical information. In addition, a forced quantification of what is in essence qualitative data may come with considerable loss of information and arbitrary decisions. Finally, work by Wagenaar and colleagues as in TAN[2] underlines that reliability and validity are not exclusively quantitative criteria: they matter in qualitative data as well. For example, an eye-witness who tells the same story on different occasions is highly reliable but without information from other sources little, if anything, can be said about the validity of this eyewitness' story. Analogously, obtaining the same scores for a group of people at different occasions may indicate a high reliability but does not yet provide a quality label for the validity of the instrument used.

To conclude: anchoring narratives as a core process

"Not everything that counts can be counted, and not everything that can be counted counts", is a quote attributed to Albert Einstein. While single pieces of evidence may provide quantitative information, for establishing the chain of evidence (e.g., Figures 2 and 3) the qualitative professional judgement cannot yet be replaced by a machine or calculator. This does not mean that quantitative information is not useful. In contrast, it appears impossible to imagine a practice of science or assessment practice where measurement and numbers have no place. Mathematics, psychometrics, statistics, and the like provide essentially very powerful tools for the mathematical modelling of empirical phenomena, be it for group-based empirical research or for the competence assessment of individuals. However, no single number makes sense in isolation or absence of supportive information. An exam score may provide partial information on some aspect of competence but needs to be evaluated in the light of other relevant behaviour of the candidate that is assessed. A 'very low' p-value (e.g., $p < 0.001$) in an experiment on a comparison between a treatment and control condition has to be evaluated in the context (e.g., theoretical framework, methodological choices, participant characteristics, nature of the data at hand) in which a study has been carried out, and even then it only starts to make sense when considering other relevant research on the phenomenon, following up with replication studies, and carrying out meta-analyses and systematic reviews.

The concept of meta-analysis provides another example of how important qualitative professional judgement is even in an exercise that at first seems to be entirely quantitative (e.g., effect sizes from each of the studies included in the meta-analysis). After all, decisions with regard to the

inclusion or exclusion of studies in a meta-analysis as well as concerning moderator variables cannot be made entirely based on numbers: it requires professional knowledge of the field, a closer study of candidate studies, and common sense.

Whether we are dealing with the question of the guilt or innocence of a suspect in a criminal case, with a question that calls for group-based empirical study or whether we wish to assess a resident's competence, qualitative and quantitative pieces of evidence have to be anchored as narratives into a chain of evidence with regard to a question of interest. As such, any kind of qualitative-quantitative divide, such as we have seen in the field of medical education, falls short and should therefore be abandoned.

Conflict of interest

The author has no conflict of interest to declare.

Author's contribution

The author testifies that he qualifies for authorship and has checked the article for plagiarism. He conceived, designed, and carried out the literature study, and wrote the full manuscript as well as the revised version of the manuscript addressing the reviewers' issues raised with the initial version of the manuscript. The author critically reviewed and approved both the initial manuscript (sent out for review) and the revised version of the manuscript (in which the reviewers' issues raised with the initial version of the manuscript have been addressed). The author is responsible for the content and similarity index of the manuscript.

References

1. Yudkowsky R, Otaki J, Lowenstein T, Riddle J, Nishigori H, Bordage G. A hypothesis-driven physical examination learning and assessment procedure for medical students: initial validity evidence. **Med Educ 2009**; 43: 729–740.
2. Wagenaar WA, Van Koppen PJ, Crombag HFM. *Anchored narratives: the psychology of criminal evidence*. New York: Palgrave Macmillan; 1994.
3. Leppink J, Pérez-Fuster P. What is science without replication? **Perspect Med Educ 2016**; 5: 320–322. http://dx.doi.org/10.1007/s40037-016-0307-z.
4. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, Van Tartwijk J. A model for programmatic assessment fit for purpose. **Med Teach 2012**; 34: 205–214.
5. Leppink J, Van Gog T, Paas F, Sweller J, Cognitive load theory: researching and planning teaching to maximise learning. In: Cleland J, Durning SJ, editors. *Researching medical education*. Chichester: Wiley & Blackwell. pp. 207–218.
6. Heeneman S, Schut S, Donkers J, Van der Vleuten CPM, Muijtjens AMM. Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. **Med Teach 2016**; 39: 44–54. http://dx.doi.org/10.1080/0142159X.2016.1230183.
7. Kane MT. Validation. In: Brennan RL, editor. *Educational measurement*. 4th ed. Westport: Praeger; 2006. pp. 17–64.
8. Messick S. Validity. In: Linn RL, editor. *Educational measurement*. 3rd ed. New York: American Council on Education/Macmillan; 1989. pp. 13–103.
9. Govaerts MJB, Van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. **Med Educ 2013**; 47: 1164–1174.
10. Govaerts MJB, Van de Wiel MWJ, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. **Adv Health Sci Educ 2013**; 18: 375–396.
11. Teunissen PW. Experience, trajectories, and reifications: an emerging framework of practice-based learning in healthcare workplaces. **Adv Health Sci Educ 2015**; 20: 843–856.
12. Stalmeijer RE, Dolmans DHJM, Wolfhagen IHAP, Muijtjens AMM, Scherpbier AJJA. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. **Acad Med 2010**; 85: 1732–1738.
13. Spruijt A, Leppink J, Wolfhagen IHAP, Bok H, Mainhard T, Scherpbier AJJA, Van Beukelen P, Jaarsma D. Factors influencing seminar learning and academic achievement. **J Vet Educ 2015**; 42: 1–12.
14. Jippes M, Driessen EW, Broers NJ, Majoor GD, Gijselaers WH, Van der Vleuten CPM. A medical school's organizational readiness for curriculum change (MORC): development and validation of a questionnaire. **Acad Med 2013**; 88: 1346–1356.
15. Kamp RJA, Dolmans DHJM, Van Berkel HJM, Schmidt HG. Can students adequately evaluate the activities of their peers in PBL? **Med Teach 2011**; 33: 145–150.
16. Chan EKH. Standards and guidelines for validation practices: development and evaluation of measurement instruments. In: Zumbo BD, Chan EKH, editors, *Validity and validation in social, behavioral, and health sciences. Social indicator research series*, vol. 54, pp. 9–24, http://dx.doi.org/10.1007/978-3-319-07794-9_2.
17. Gehlbach H, Artino AR, Durning SJ. AM last page: survey development guidance for medical education researchers. **Acad Med 2010**; 85: 925.
18. Artino AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE guide no. 87. **Med Teach 2014**; 36: 463–474.
19. Kline RB. *Principle and practice of structural equation modeling*. 3rd ed. London: The Guilford Press; 2010.
20. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Newbury Park, CA: Sage; 1991.
21. McCutcheon AL. *Latent class analysis. Quant applications in the social sciences series no. 64*. Thousand Oaks, CA: Sage; 1987.
22. Lazarsfeld PF, Henry NW. *Latent structure analysis*. Boston: Houghton Mill; 1968.
23. Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. **Adv Health Sci Educ 2015**; 20: 829–834.
24. Lafleur A, Côté L, Leppink J. Influences of OSCE design on students' diagnostic reasoning. **Med Educ 2015**; 49: 203–214.
25. Picho K, Maggio LA, Artino AR. Science: the slow march of accumulating evidence. **Perspect Med Educ 2016**; 5. http://dx.doi.org/10.1007/s40037-016-0305-1.

How to cite this article: Leppink J. Evaluating the strength of evidence in research and education: The theory of anchored narratives. J Taibah Univ Med Sc 2017;12(4):284–290.