International Journal of Artificial Intelligence in Education (2020) 30:205–235
https://doi.org/10.1007/s40593-019-00188-w

ARTICLE

Impact of an Artificial Intelligence Research Frame on the Perceived Credibility of Educational Research Evidence

Mutlu Cukurova[1] · Rosemary Luckin[1] · Carmel Kent[1]

Published online: 18 December 2019

Abstract

Artificial Intelligence (AI) is attracting a great deal of attention and it is important to investigate the public perceptions of AI and their impact on the perceived credibility of research evidence. In the literature, there is evidence that people overweight research evidence when framed in neuroscience findings. In this paper, we present the findings of the first investigation of the impact of an AI frame on the perceived credibility of educational research evidence. In an experimental study, we allocated 605 participants including educators to one of three conditions in which the same educational research evidence was framed within one of: AI, neuroscience, or educational psychology. The results demonstrate that when educational research evidence is framed within AI research, it is considered as less credible in comparison to when it is framed instead within neuroscience or educational psychology. The effect is still evident when the subjects' familiarity with the framing discipline is controlled for. Furthermore, our results indicate that the general public perceives AI to be: less helpful in assisting us to understand how children learn, lacking in adherence to scientific methods, and to be less prestigious compared to neuroscience and educational psychology. Considering the increased use of AI technologies in Educational settings, we argue that there should be significant attempts to recover the public image of AI being less scientifically robust and less prestigious than educational psychology and neuroscience. We conclude the article suggesting that AI in Education community should attempt to be more actively engaged with key stakeholders of AI and Education to help mitigate such effects.

✉ Mutlu Cukurova
m.cukurova@ucl.ac.uk

Rosemary Luckin
r.luckin@ucl.ac.uk

Carmel Kent
k.carmel@ucl.ac.uk

[1] UCL Knowledge Lab, University College London, 23- 29 Emerald Street, London WC1N 3QS, UK

Keywords AI in Education · Perceived credibility · Research evidence · The in-credible AI effect

Introduction

Within the last decade, the private sector has invested heavily in large scale projects to develop AI that can interact with humans (Hall and Pesenti 2017). This has resulted in many of these technologies becoming part of our everyday practice in work, home, leisure, healthcare, social care, and education. There is a great benefit to be gained from some applications of AI, as they have the potential to save time, reduce the human effort to perform tasks and reduce costs (Yang et al. 2017). Through the automation of repetitive and monotonous tasks, AI systems can relieve humans from both dangerous and menial tasks, improve well-being through the provision of reliable care assistance for the ageing population, improve service encounters through standardisation, and provide companionship and affective aids for different user groups (Winfield and Jirotka 2018). As argued by Malone (2018) AI systems also have the potential to augment human intelligence in machine-human collaborations as 'superminds'. In healthcare for example, where deep learning algorithms have been trained to identify pathology automatically from optical coherence tomography (OCT) scans to enable early detection and triage of sight-threatening diabetic retinopathy and age-related macular degeneration, to advise doctors, and interpret fMRI scans (De Fauw et al. 2018). These technologies are developing at a rapid pace and they are increasingly entering our everyday lives. The speed of AI invention and application as well as the excitement associated with it has lead researchers to question whether AI explanations exert a "seductive allure" on individuals, leading them to judge bad explanations or arguments framed in AI more favourably (Giattino et al. 2019). A similar phenomenon was noted in the early 2000s and 2010s when the term 'neuroscience' was similarly popular and its 'seductive allure' was noted by some researchers (McCabe and Castel 2008; Im et al. 2017).

On the other hand, not all mentions of AI are associated with positive public attitudes (European commission 2015, BSA 2015), and concerns regarding the potentially harmful impact of AI technologies are also often raised in the media and public rhetoric. Moreover, highly respected academics and public figures have exacerbated the construction of dystopian scenarios of AI machines causing existential consequences to humankind. For instance, Steven Hawking warns that AI could end mankind[1] and Elon Musk has argued that AI is highly likely to destroy humans.[2] Besides, there are other arguments on the observed and expected impact of AI on the future of the workforce, and related fears around mass unemployment (Frey and Osborne 2013; Brynjolfsson and Mcafee 2014). These negative annotations have the potential to skew public perceptions and avoid or ignore AI systems. For instance, although evidence-based algorithms more accurately predict the future than do human forecasters, when forecasters are deciding whether to use a human forecaster or an algorithm, they often choose the human forecaster (Dietvorst et al. 2015).

[1] https://www.bbc.co.uk/news/technology-30290540

[2] https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-artificial-intelligence-openai-neuralink-ai-warning-a8074821.html

These recent concerns lend considerable weight to the need to explore the public's perceptions of AI and its potential impact on the credibility judgements of research evidence when in comparison to other areas of science. Investigating the public perceptions of AI is important as it can lead to regulatory activity with potentially serious repercussions in society (Stone et al. 2016) as well as for helping us define social policy (Manikonda and Kambhampati 2018). More specifically for AI in Education, potential aversion of the public to AI is costly for society at large. For instance, AI in Education systems can be comparable to human tutors in terms of their effectiveness (i.e, VanLehn 2011; du Boulay 2016), however many people remain resistant to using them, and actively demonstrate against their use in schools.[3] For effective adoption of AI, it is important to create a shared understanding between the key stakeholders of AI Technologies including public, educators, and academia (Cukurova et al. 2019)

In this article, we explored the public perceptions of AI and their perceived credibility of research evidence about education framed to different areas of scientific research, specifically: educational psychology, neuroscience, and AI. While in the past, there have been some attempts at investigating the public perceptions of AI through opinion polls (Gaines-Ross 2016); with a longitudinal study of articles (Fast and Horvitz 2016), and through social media analysis (Manikonda and Kambhampati 2018); to the best of our knowledge, this is the first investigation of whether our attitudes towards AI blur our perceptions of the credibility of educational research evidence.

Public Perceptions of AI and the Credibility of Educational Research Evidence

Studying public perceptions of AI is quite challenging, not the least due to the lack of a clear definition of the term AI. As argued by Stone et al. (2016), laypeople and experts alike have varied understandings of what AI means. Even in the narrowest, engineering-centric definitions, AI refers to a broad constellation of computing technologies including a broad range of machine learning approaches (Fast and Horvitz 2017). There are a few previous studies looked at the impact of AI on public perceptions as well as the range of future outcomes (Dietterich and Horvitz 2015). For instance, Fast and Horvitz (2016) conducted a longitudinal study of articles published on AI in the New York Times between January 1986 and May 2016. This study revealed that from 2009 the discussion on AI has sharply increased and is more optimistic than pessimistic. Gaines-Ross (2016) investigated the lay people's perception of AI and found out that those individuals who do not have a professional background in technology generally have positive perceptions of AI. More recently, Manikonda and Kambhampati (2018) collected and analysed over two million AI-related tweets posted by over 40,000 people and showed that 1) the sentiments expressed in the AI discourse are more positive than an average twitter discourse 2) lay public tend to be more positive about AI than expert tweeters and 3) women tend to be more positive about AI impacts than men. In general, existing research on the public

[3] https://www.nytimes.com/2019/04/21/technology/silicon-valley-kansas-schools.html

perceptions of AI shows greater levels of optimism than pessimism about AI (60 Minutes, 2016), but they also show increasing existential fear and worry about jobs (BSA 2015). None of the existing studies compares AI with other subject areas of Educational Psychology and Neuroscience, our study presented here is the first to compare these three subject areas related to educational contexts and show the impact of public perceptions of AI on their judgment of the educational research credibility.

Credibility is described as the *believability* of a source or a message (Metzger and Flanagin 2015). Here, it is important to clarify the difference between *veracity* and *credibility*. Veracity is a concept that refers to the true or false nature of a piece of evidence (Derczynski et al. 2017). On the other hand, credibility is a perceptual variable which is subjectively perceived by recipients and is not an objective attribute (Shariff et al. 2017). Wassmer and Eastman (2005) differentiate between actual and perceived credibility, whereby actual credibility can be equalled with veracity. Here, we are investigating the *perceived credibility* of research evidence presented in different subject frames. It is important to note that in this study, the evidence presented in all subject frames is the same, and therefore the veracity of the evidence is also the same (please see the first two columns of appendix C). However, we investigate the extent to which the *perceived credibility* of the evidence differs.

The presentation of unfamiliar information has been demonstrated to impact upon peoples' judgment of the credibility of the evidence presented in detail (Appelman and Sundar 2016). Perceived credibility, which is a subjective measure shaped by the impressions of the observer, can be formed through two distinct forms of information processing. It can either be shaped through a central route or a peripheral route (the Heuristic Systematic Model, Chaiken 1987). In the central route, receivers have higher cognitive processing ability and/or are highly motivated, which increases the possibility of them to be more engaged with and to scrutinise a message. On the other hand, the peripheral route is taken when people are neither motivated nor able to cognitively process the information provided. In this route, therefore, the perceived credibility is shaped by peripheral cues or heuristic rules, the subject frame in which the evidence is presented, for example. Similar ideas are echoed in other models (the elaboration likelihood model of persuasion, Petty and Cacioppo 1986). It can be argued that unfamiliar research evidence is often processed in a more peripheral way rather than a central way.

Heuristic judgements are found to be triggered by specific aspects of the information, source, or interaction situation (Kruglanski and Gigerenzer 2011). Here, we compare three subject discipline framings for the same research evidence: Educational Psychology, Neuroscience, and AI. In the case of neuroscience, suggestions to improve educational practice with neuroscience findings have been explained as an appeal to neuroscience findings rather than the actual contribution of those findings (Hardiman et al. 2012). This, in turn, affects the perceived credibility of educational neuroscience as a scientific discipline (Schwartz et al. 2012). In the case of AI, to the best of our knowledge, such an exploration of the potential framing impact of AI on research findings in Education has not been investigated previously.

The fact that perceived credibility is often processed by readers with peripheral cues and heuristic rules and in recognition of the negative images and science fiction associations of AI in the media we hypothesise that:

i) AI in Education is perceived as less relevant to learning, adheres less to the scientific methods, and less prestigious than Educational Psychology and Neuroscience;
ii) research evidence framed in AI in Education is considered as less credible than the same research evidence framed in Educational Psychology which, in turn, is considered as less credible than Neuroscience frame due to 'the seductive allure of neuroscience' (Im et al. 2017).

In this paper, we test these hypotheses and present our results about whether the AI framing of research evidence impacts on the perceived credibility of research evidence in Education.

Literature Review

The perceived credibility of AI framing has not yet been studied. However, the relevant phenomenon of 'the seductive allure of neuroscience' is widely studied and reported in the literature (McCabe and Castel 2008; Weisberg et al. 2008; Im et al. 2017) which will be reviewed in this section.

The Seductive Allure of Neuroscience

People have been shown to give more weight to evidence framed within neuroscience. Weisberg et al. (2008) asked participants to judge the quality of arguments in articles on psychological phenomena, and their results show that the inclusion of a neuroscience finding reduces the ability of novice participants to distinguish the good explanations from the bad ones. Similarly, McCabe and Castel (2008) gave their participants a one-page summary of a cognitive neuroscience finding written for the popular press in the control condition. In the experimental condition, participants were provided with the same information accompanied by a bar chart or a brain image from an fMRI scan. Participants rated the scientific reasoning most highly when the information was accompanied by an fMRI image, so the authors concluded that neuroscience explanations are more credible when they are accompanied by brain images. Nevertheless, more recent studies, failed to replicate these findings. Farah and Hook (2013) used a similar strategy to present neuroscience explanations with a brain image, with a bar chart, or on its own. They found that the addition of a brain image did little to change the perceived quality of scientific explanations. Similarly, Michael et al. (2013) undertook a comprehensive study on the topic with almost two thousand participants and reached similar conclusions to Farah and Hook (2013). More recently, to investigate the impact of superfluous neuroscience explanations, Fernandez-Duque et al. (2015) undertook four experiments with 385 college students. Students were asked to read brief descriptions of psychological phenomena, each one accompanied by an explanation of varying quality and followed by superfluous information of various types. The authors concluded that superfluous neuroscience information increased the judged quality of the argument for both good and bad explanations, whereas accompanying fMRI pictures had no impact above and beyond the neuroscience text. Although, recent evidence shows that the addition of a brain picture does little to increase the perceived quality of a neuroscientific explanation (Hook and Farah 2013; Michael et al. 2013; Schweitzer et al. 2013; Gruber and Dickerson 2012); these studies do not investigate whether extraneous neuroscience information (either pictorial or textual) has an influence on the interpretations of research evidence that is non-neuroscientific in all other aspects. This is an important distinction. It leads to the suggestion that, rather than perceptual issues of brain images, the seductive allure of neuroscience might be driven by the conceptual properties of brain-related subjects. In order to investigate whether the presence of neuroscience information exerts an undue influence on judgments of research quality, Fernandez-Duque et al. (2015) undertook a series of experiments and concluded that superfluous neuroscience information was more alluring than social science information and more alluring than information from prestigious hard sciences such as chemistry, biology, and genetics.

This finding illustrates that neuroscience bias might be conceptual rather than pictorial. Such bias may exert undue influence on judgments of evidence credibility, as has been supported more recently by evidence generated from various disciplines. More specifically, within the context of Education, Im et al. (2017), recruited 320 participants from the general public and asked them to judge the credibility of research articles that are framed with neuroscientific verbal, graphical, and brain images input. Their results showed that members of the public judge the credibility of educational articles that have extraneous verbal neuroscience findings and brain images higher than those articles with no extraneous neuroscience information. Moreover, the effect persists even after controlling for individual differences in familiarity with education, attitudes towards psychology and knowledge of neuroscience. Similarly, the seductive allure of neuroscience has been observed to different extents in different fields including law (Schweitzer and Saks 2011), and marketing (Ariely and Berns 2010).

Possible Explanations of the Seductive Allure of Neuroscience and Their Relevance to the Incredible AI Effect

There are various hypotheses generated to explain the potential effect of the seductive allure of neuroscience. The most common one is the perceptual processing hypothesis (Keehner et al. 2011; Reber & Schwarz, 1999) which argues that perceptual features of brain pictures, such as their three-dimensionality, biases people's judgement of the accompanying text. These arguments can be likened to some of the examples of AI in education. These examples include AI texts accompanied by pictorial information ranging from accurately represented algorithms to completely irrelevant representations of futuristic robots. Another hypothesis is that the prestige of neuroscience as a hard science is higher than some other sciences and that this prestige biases people's judgment of the research evidence (Keil et al. 2010). Although Fernandez-Duque et al. (2015) show that the prestige of "hard" sciences does not extend to natural sciences such as biology, chemistry, and genetics, it might well extend to brain-related subjects, such as neuroscience and AI. Explanations that invoke neuroscience may be viewed as reflecting greater expertise, and similarly, explanations that invoke AI may also be considered to reflect greater expertise. Similarly, it might also be the case that the jargon of certain disciplines might create an impression that the information presented is more or less credible. For instance, just as adding random nonsensical mathematical equations to abstracts increases their perceived credibility (Eriksson

2012), using the jargon of AI and algorithms might, therefore, affect the perceived credibility of the presented information. It was also argued that neuroscience's role as the "engine of the mind", in the sense that brain models presented in neuroscience are the best explanations of how the mind works, might convince people that the information framed in neuroscience is more credible (Fernandez-Duque et al. 2015). This explanation also aligns with the findings that the allure of neuroscience does not extend to the prestige of the "hard sciences", but the information should be somehow framed in neuroscientific explanation. On the other hand, the authors' brain-as-engine-of-mind hypothesis, could equally relate to AI, because most of the models and explanations in AI are intertwined and influenced by each other and a better understanding of neuroscience plays a vital role in building AI (Hassabis et al. 2017).

Limitations of Previous Studies and Confounding Factors in Credibility of Research Evidence Evaluations

Most aforementioned existing studies are criticised for various limitations. For instance, earlier studies by Weisberg et al. (2008) and McCabe and Castel (2008) were criticised for the amount of information presented in control and experiment conditions, because these were not equal. It is, therefore, possible that the addition of the neuroscience information simply acted to conceal the circularity of the explanations. Surprisingly, the same limitation concerning the article length confounding with the results was also present in more recent studies (Im et al. 2017). However, some other studies showed that even when article length is equated, the seductive allure of neuroscience effect remains (Weisberg et al. 2015).

The earlier neuroscience studies also failed to investigate whether the results were associated with other individual differences among the participants, such as familiarity with the subject, attitudes towards the subjects, particular demographic features of the participants, their prior knowledge of the topic, or their reasoning ability. These are important confounding factors in the investigation of the AI context. For instance, Fernandez-Duque et al. 2015 found that the ability to reason with analytical thinking, and, a belief in dualism or free will do not protect a participant from the neuroscience bias in judging the credibility of research evidence. However, people do rate the credibility of research higher when the findings are consistent with their prior belief (Scurich and Shniderman 2014). Additionally, when the participants have limited knowledge of the scientific method being reported in what they read, the neuroscience bias of the credibility they award to the research is bigger (Rhodes et al. 2014). Therefore, the familiarity of the participants with the topic is a significant confounding factor that should be taken into account (Im et al. 2017).

Methodology

In this study, we have two hypotheses as presented earlier and we investigate them with three research questions. The first research question is whether the framing discipline of the same educational research evidence has an impact on the public's perceived credibility of it. We investigate three disciplinary frames: neuroscience, educational psychology and AI. Due to divergence between the realities of AI and the ways that it is portrayed in various forms of media, we hypothesise that people's credibility value will be skewed compared to other articles that present the same evidence in educational psychology and neuroscience frames. This is the first investigation of the concept, we have therefore not brought in any experimental levelling to change the amount or type of framing presented to the participants.

Our second research question is to investigate how a potentially skewed public perception of evidence framed through AI would compare to the seductive allure of neuroscience effect.

To investigate these two research questions, we use similar articles to those used in a recent study investigating the seductive allure of neuroscience effect in education (Im et al. 2017). Then, we compare the credibility values of neuroscience and AI frames with the credibility values of an Educational Psychology frame for the same research evidence.

Our last research question is about the attitudes of the public towards Educational Psychology, Neuroscience, and AI and how some of the individual differences impact on the credibility judgments of the public. Based on the literature reviewed above, we are particularly interested in two potential confounding variables: whether participants' familiarity with, and their attitudes towards the fields of Educational Psychology, Neuroscience, and AI account for any particular bias in their judgments of research evidence credibility.

Participants, Design, and Study materials

We target general public participants, and we, therefore, recruited participants via an online survey development cloud-based software company. This is a platform where adult 'workers' from across the world sign up to perform online 'jobs' for compensation (Follmer et al. 2017). The online survey was made available to members of the public who live in the United Kingdom and the United States, who are over =18-year-old, and whose primary language is English. The survey was introduced within the context of asking participants for their opinion on a variety of short educational articles and their evaluation of the amount of credibility they assign to each article. The participants were informed that the survey would take around 15–25 min of their time. The survey was anonymous and no personal details were required, and participants had the right to withdraw from the survey at any point. Each participant was compensated with £2.25 upon completion of the survey. We used the automated block randomiser functionality of the online survey platform to randomly assign participants in to one of the disciplinary frames of AI, Educational Psychology, or Neuroscience. We also used item randomisation for credibility and attitude surveys to avoid item order bias. Each participant gave informed consent through an online confirmation.

605 respondents filled in our questionnaire, out of which only 502 provided a full response. After cleaning responses indicative of random clicking, 345 respondents remained. 157 participants (31%) were excluded because they failed the attention check items designed to catch participants who do not engage with the survey but instead respond randomly (Oppenheimer et al. 2009). This exclusion rate is comparable to those other studies which recruited online participants (i.e. Chandler et al. 2014; Im et al. 2017). Each of the participants was randomly assigned to a different version of the questionnaire, containing article excerpts presenting the same evidence in the three different subject frames (see Table 1 below). For details on the data collection tools used please see the appendix provided.

Table 1 The three experiment groups

Group	Number of subjects	Percentage
1: Neuroscience	106	30.7
2: Educational psychology	121	35.1
3: Artificial intelligence	118	34.2

A set of four background questions and nine attitude questions was shared with all the participants, while the rest of the questionnaire contained three discipline-specific articles assigned differently to each group. For each of the three articles, the participants were asked to rate five different credibility scores and one score assessing their familiarity with the topic. The frequency charts of the four background questions, regarding gender, education level, their major, and age, are shown in the Fig. 1.

These demographics indicate that the majority of the sample was aged between 25 and 34, had studied to tertiary level, approximately two-thirds were female, and most of them were non-specialists in the fields of AI, Educational Psychology, and Neuroscience. More specifically, the participants involved 66 educators, 43 computer scientists, 40 psychologists, 6 neuroscientists, and 190 participants from other professional areas. 130 participants had a high school degree, 112 had an undergraduate, 66 had a Masters' degree and 32 had a PhD. The nine attitude items contained three questions examining the attitudes of the participant to each of the three disciplines. Specifically, all participants were asked about (1) whether the discipline can help understand how children learn; (2) whether researchers practising in the discipline adhere to scientific methods; and (3) whether society views the discipline as prestigious - for each of the three disciplines.

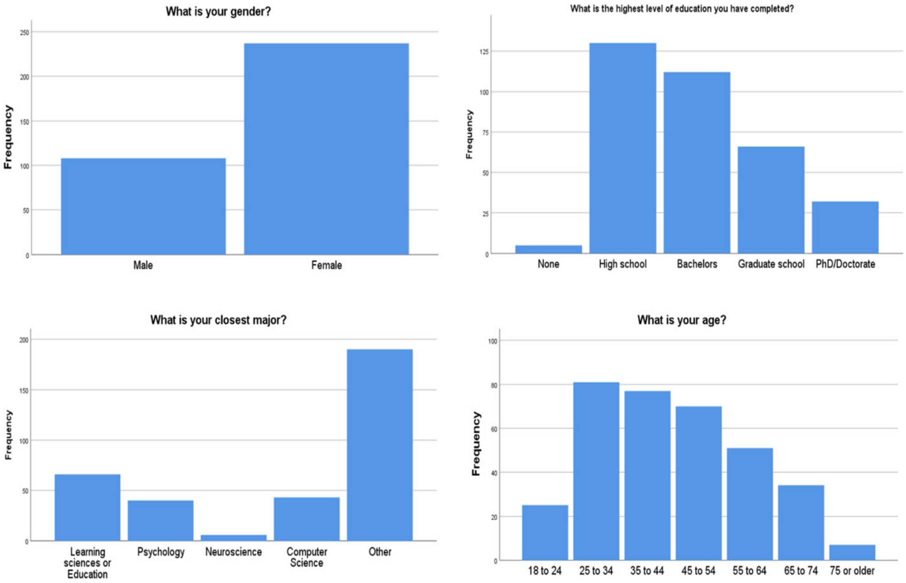


Fig. 1 Frequency of the participants' demographic feautures

The dependant construct of credibility was operationalized by five different variables (all measured using 7 points Likert scale, 7 = "*Strongly agree*" to 1 = "*Strongly disagree*"). The five Likert items each participant scored for each of the three articles they were exposed to are summarized in Table 2 along with the variable name we will use in this paper.

We considered perceived credibility as a multidimensional construct, which mainly relies on perceptions of scientific argument, empirical evidence, agreeable conclusions, helping to understand, and well-written natures of the articles presented. Multidimensional scales assessing different aspects of perceived credibility are considered as superior measurement tools than single item measurements (Metzger and Flanagin 2015). The unidimensionality of the five items was confirmed by a principal components analysis (Hattie 1985) in a previous study in which the first principal component had an eigenvalue of 4.125 and accounted for 82.5% of the variance. None of the remaining principal components had an eigenvalue >1 or accounted for more than 8.1% of the variance (Im et al. 2017). Hence, for statistical analysis, we have averaged the scores (Uebersax 2006) of each of the five Likert items throughout the three articles each subject scored. The resulting five averaged scores were moderately negatively skewed. Therefore, to make them near normal, we have transformed and reflected them using the transformation suggested by Tabachnick and Fidell (2007): $\sqrt{Largest\ score + 1 - score}$.

After the transformation, three data points were shown to be more than 1.5 box-lengths from the edge of their box with regard to the AGREEABLE variable. Looking closer into those outliers, we have decided to remove them since these participants gave repetitive answers of either 7 s or 1 s Likert scores. However, looking for multivariate outliers, we found eight observations with Mahalanobis distance larger than the chi-square critical variable of 20.52 (for five degrees of freedom and $p < .001$). As these all looked like genuine data points, we have decided not to remove those.

Results

The Public Attitudes towards Three Disciplines

A Friedman test was run to determine if there were differences in the subjects' views on how the three disciplines help us understand how children learn, on how the researchers

Table 2 The five dependent variables, their alias names as used throughout this paper and their reported MANCOVA weights

Likert item	Variable name used here	MANCOVA weights
The article was well written	WELL-WRITTEN	1.186
The article helped me understand the topic	HELPS-UNDERSTANDING	4.445
The scientific argument in the article made sense	SCIENTIFIC	−0.003
The article offered strong empirical evidence for its conclusions	EMPIRICAL	−4.166
Do you agree or disagree with the conclusions of the article?	AGREEABLE	−0.295

in the discipline adhere to scientific methods and on how society views the disciplines as prestigious. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. The attitudes on all three aspects were found to be significantly different between the disciplines.

$\chi^2(2) = 142.623$, $p = 0.000$ for the helps us understand how children learn item. Post hoc analysis revealed statistically significant differences between AI (median = 5) and EDPSY (median = 6) ($p = 0.000$) and between AI and NS (median = 6) ($p = 0.000$).

$\chi^2(2) = 82.183$, $p = 0.000$ for the adhering to scientific methods item. Post hoc analysis revealed statistically significantly different between AI (median = 5) and EDPSY (median = 6) (p = 0.000), between AI and NS (median = 6) (p = 0.000) and between NS to EDPSY in favor of NS ($p = 0.047$).

$\chi^2(2) = 54.072$, $p = 0.000$ for the prestigious viewed by society item. Post hoc analysis revealed statistically significantly different between AI (median = 5) and NS (median = 6) ($p = 0.000$) and between NS to EDPSY (median = 5) ($p = 0.000$).

The post-hoc tests reveal the inferiority of the AI discipline frame and the slight superiority of the neuroscience discipline frame, as is shown in Fig. 2.

The Seductive Allure of a Subject in Research Credibility

The credibility dependant construct was operationalized by five different averaged scores, and therefore we ran a one-way MANCOVA to test for the differences between the three experimental groups, while controlling for the subjects' familiarity with the discipline. There was an approximately linear relationship between each pair of the dependent variables, as well as between the covariate, familiarity with the topic, and the dependent variables in each group, as assessed by scatterplots. There was homogeneity of regression slopes, as assessed by the interaction term between familiarity average and group, $F(10, 670) = 1.539$, $p = 0.121$. The assumption of the equality of variance-covariance matrices was violated, as Box's test of Equality of Covariance Matrices was shown to be statistically significant ($p < .001$). Since the groups have similar and large sizes, this violation did not justify not running the MANCOVA. However, we decided to make a conclusion based on both the Wilk's Lambda MANOVA test, and the Pillai's Trace test, which is more robust to this data violation.

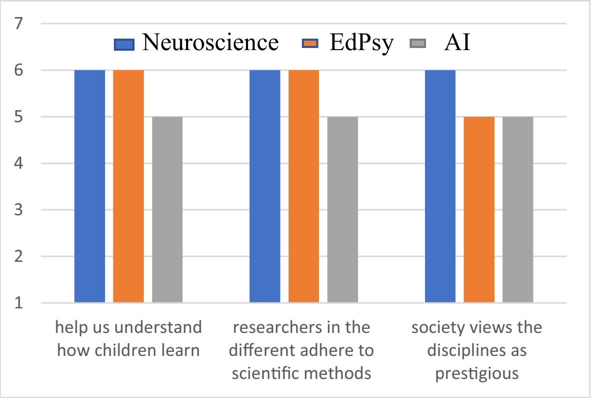


Fig. 2 Median Likert scores of the participants' attitudes

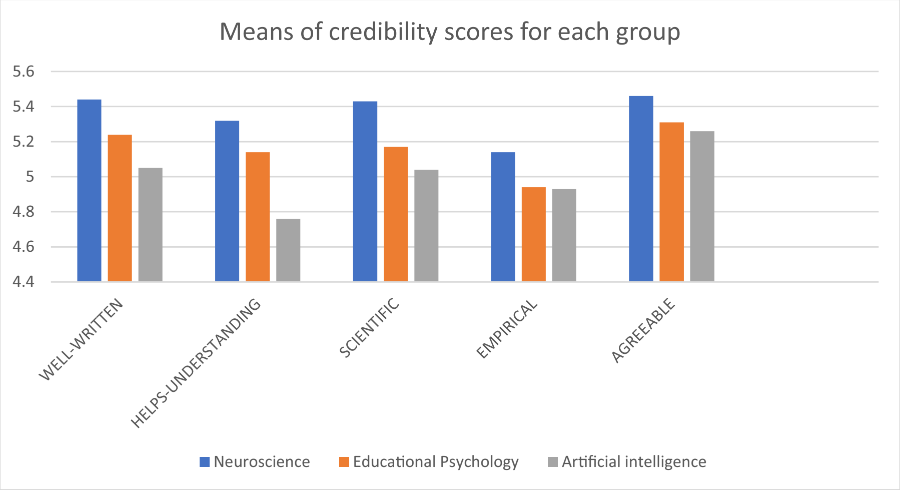
There was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance ($p > .05$) for all variables. Residuals were normally distributed, as assessed by Shapiro-Wilk's test ($p > .0033$ using a Bonferroni adjustment). There was no multicollinearity detected by both Pearson and Spearman tests: all dependent variables were significantly moderately (<0.9) correlated. Table 3 below summarises the mean values and standard deviations for each subject frame and credibility dimension.

The MANCOVA resulted in a statistically significant difference between the groups on the combined dependent variables, after controlling for the participants' discipline familiarity, F(10, 674) = 2.488, $p = 0.006$; Wilks' $\Lambda = 0.930$; partial η2 = 0.036. Pillai's Trace test has also shown a statistically significant difference, F(10, 676) = 2.484, $p = 0.006$; Pillai's Trace = 0.071; partial η2 = 0.035. The multivariate partial η2 indicates that approximately 3.5% of the multivariate variance of the dependent variables is associated with the difference between the groups. This is a small to medium effect size and the relatively small size of the effect might be accounted for the combination and transformation of the variables.

As a follow up to the MANCOVA, none of the univariate ANCOVA tests was found to be significant at the 0.01 adjusted alpha level (while protecting against type I error 0.05 divided by five). Since all the five variables operationalize together to the same theoretical construct of credibility (Im et al. 2017), we decided to follow-up the significant MANCOVA result, analysing the difference between the groups in the newly created multivariate composite variable. The weights assigned by the MANCOVA to each of the five independent scores are shown in the rightmost column of Table 2. As can be seen, the

Table 3 Unadjusted means and standard deviation (in brackets) for each subject frame and each dependent variable

	Well-written	Helps-understanding	Scientific	Empirical	Agreeable
Neuroscience	5.44 (0.13)	5.32 (0.13)	5.43 (0.13)	5.14 (0.14)	5.46 (0.12)
Educational Psychology	5.24 (0.11)	5.14 (0.12)	5.17 (0.10)	4.94 (0.12)	5.31 (0.11)
Artificial intelligence	5.05 (0.12)	4.76 (0.14)	5.04 (0.13)	4.93 (0.13)	5.26 (0.10)



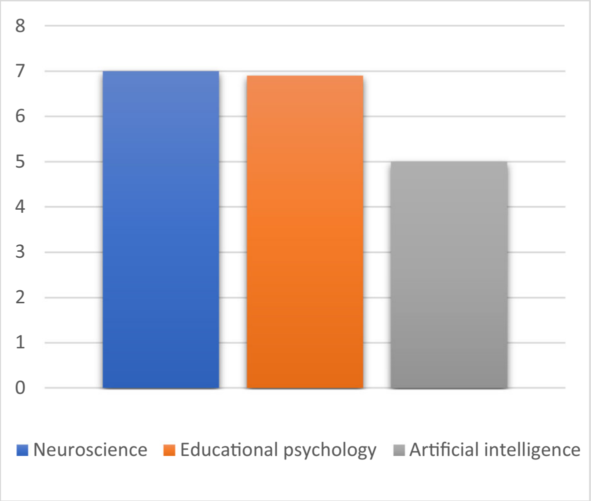


Fig. 3 Canonical credibility mean values for all participants

most influential variables (in two different directions) are HELPS-UNDERSTANDING and EMPIRICAL.

Next, we used these weights to compute the composite credibility variable for each participant. A Kruskal-Wallis H test was run to determine the differences in the composite credibility score between the three groups. Distributions of the credibility score were similar for all groups, as assessed by visual inspection of a boxplot. Median credibility scores were statistically significantly different between groups, $H(2) = 18.463$, $p = 0.000$.

Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. The post-hoc analysis revealed statistically significant differences in median scores between the Neuroscience and AI ($p = .001$) disciplines, and between the Educational psychology discipline and AI ($p = .001$). Figure 3 shows the means of the composite credibility scores after being untransformed and re-reflected to reverse the initial transformation (Tabachnick and Fidell 2007).

Effects of expertise and attitudes towards the subjects on the results

To determine which of the demographic background and attitude variables has a significant effect on the credibility, we ran six multiple linear regressions. One regression model to find the independent variables which significantly contribute to the newly created composite credibility score, and then another five regression models for each of the univariate credibility scores. All multiple regression models significantly predicted the different six credibility scores. Interestingly, as summarized in the table below, three main factors repetitively contributed to high credibility scores were: (1) the attitude that Educational psychology researchers adhere to scientific methods (in yellow in the below table); (2) the attitude that society views Educational psychology as a prestigious discipline (green); and (3) the attitude that AI research can help us understand how children learn (brown). Moreover, only (1) was a significant predictor of the composite credibility score, and participants age was a significant factor for help-understanding and empirical dimensions.

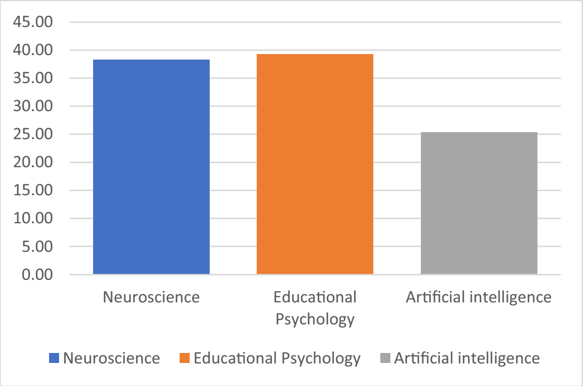


Fig. 4 Canonical credibility mean ranks for educators

The Negative Effect of AI frame in Research Credibility Judgements of Educators

Although, our multinomial regression results did not show participants' major is a factor contributing to the main effect observed in our study; due to our particular research interest in educators' credibility judgments, and to the central role of educators in the preparation of AI perceptions within the younger generations, we investigated whether similar results were obtained for educators. For this purpose, a Kruskal-Wallis H test was run to determine the differences in the composite credibility score between the three groups within just those majoring in education. The number of educators in each condition was as follows: the neuroscience condition ($n = 20$), the educational psychology (n = 20), and the AI condition ($n = 26$). Distributions of the credibility score were not similar for all groups, as assessed by visual inspection of a boxplot, therefore we could not compare median scores, but we compared the mean ranks. The distributions of composite credibility scores were statistically significantly different between the groups, H(2) = 7.535, $p = .021$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post hoc analysis revealed statistically significant differences in the credibility scores' mean ranks between the AI condition (25.37) and the EdPsy condition (39.28) ($p = .044$), but not between the other post hoc pairwise comparisons (Fig. 4).

Discussion

In this study, we investigated whether different disciplinary frames for the same educational research evidence affect the general public's perceptions of the credibility of that research evidence. In this section, we will discuss the results presented in the previous section, consider their implications and the current study's limitations. We conclude the section with a discussion of future research directions and some suggestions to the AIED community.

Going back to our first research question: Is the perceived credibility of evidence affected by the disciplinary field within which it is framed? The primary finding was that the perceived credibility of educational research evidence decreases when it is framed in AI. To the best of our knowledge this is the first study investigating this impact which we name here as "*the in-credible AI effect*". The potential impact of subject framing on the credibility

Table 4 Multinomial regression results

Dependant variable	Predictor variables	B re-reflected	Std. Error of coefficient	Standardized coefficient (not re-reflected)	F Value	Adj. R^2
Composite credibility					F(15, 329) = 2.189*	0.049
	Educational psychology researchers adhere to scientific methods.	0.531*	0.228	−204		
Helps-understanding					F(15, 329) = 15.567**	0.388
	What is your age?	(−0.028)*	0.012	0.103		
	Educational psychology researchers adhere to scientific methods.	0.078**	0.02	−0.271		
	Society views Educational psychology as a prestigious discipline.	0.06**	0.014	−0.231		
	Artificial Intelligence research can help us understand how children learn.	0.034*	0.015	−0.14		
Well-written					F(15, 329) = 18.632**	0.435
	Educational psychology researchers adhere to scientific methods.	0.084**	0.019	−0.306		
	Society views Educational psychology as a prestigious discipline.	0.065**	0.013	−0.266		
Scientific					F(15, 329) = 17.212**	0.414
	Educational psychology researchers adhere to scientific methods.	0.068**	0.019	−0.248		
	Society views Educational psychology as a prestigious discipline.	0.053**	0.013	−0.216		
	Artificial Intelligence research can help us understand how children learn.	0.033*	0.014	−0.143		
Empirical					F(15, 329) = 17.133**	0.413
	What is your age?	(−0.029)*	0.012	0.105		
	Educational psychology researchers adhere to scientific methods.	0.067**	0.02	−0.232		
	Society views Educational psychology as a prestigious discipline.	0.069**	0.014	−0.267		
Agreeable					F(15, 329) = 14.391**	0.369
	Educational psychology researchers adhere to scientific methods.	0.054*	0.018	−0.214		
	Society views Educational psychology as a prestigious discipline.	0.029*	0.013	−0.127		
	Artificial Intelligence research can help us understand how children learn.	0.029*	0.014	−0.137		

* $p < 0.05$, ** $p < 0.001$

judgments of research evidence is argued to be very small in online studies (Michael et al. 2013). However, our results revealed statistically significant differences in median scores between the Neuroscience and AI ($p = .001$) disciplines, and between the Educational psychology discipline and AI ($p = .001$). On the other hand, our results show that such an effect is not observed for neuroscience framing. This result is consistent with previous work which shows that the seductive allure effect of neuroscience is driven by rich depictions of brain structure and function (McCabe and Castel 2008), yet when an article describing a psychological finding alone is contrasted with an article that additionally includes an extraneous verbal description of a neuroscience frame no effect is observed (Hook and Farah 2013; Schweitzer et al. 2013; Im et al. 2017). When we looked at the particular group of educators to see if the results we observed for the general public would also obtain for teachers, we found similar results that AI framed articles were considered statistically significantly less credible than educational psychology and neuroscience framed ones by educators.

One significant limitation of the previous studies was the confounding factor of article length. In this study, we kept the article length the same for neuroscience, AI, and the control condition of educational psychology and found that the framing of neuroscience had no statistically significant impact on the perceived credibility of the research evidence. This result is aligned with other recent studies (i.e Hook and Farah 2013; Schweitzer et al. 2013; Gruber and Dickerson 2012) and might be due to a decline in the seductive allure of neuroscience effect (Schooler 2011). Based on these results, we could also expect a decline in the “in-credible AI effect”, if and when reliable and valid scientific research and representations of AI are reflected in press coverage, and the public image of the discipline is recovered.

Our second question investigated the public perceptions of the subjects of AI, neuroscience, and educational psychology. Our questionnaire involved nine items: three questions examining the attitudes of the participant about each of the three disciplines. Specifically, all participants were asked about: (1) whether the discipline can help understanding how children learn; (2) whether researchers practising in the discipline adhere to scientific methods; and (3) whether society views the discipline as prestigious for each of the three disciplines. In terms of item one, although there was no statistically significant difference between educational psychology and neuroscience, AI was considered as less useful to help us understand how children learn. Furthermore, Neuroscience was considered as the most adherent to scientific methods, and Artificial Intelligence was considered the least adherent to scientific methods. Both differences were statistically significant between the three disciplines. On the other hand, neuroscience was considered as a more prestigious discipline than AI and Educational Psychology. These results are aligned with the previous research which shows that college students’ perceptions of neuroscience as a more prestigious discipline than natural science, psychology, social psychology, and social sciences (Fernandez-Duque et al. 2015, experiment 2). However, our research extends the results about the perceptions of college students to the general public sampled here. The same study also shows that any potential effect of the seductive allure of neuroscience is not due to the prestige of the discipline (ibid, experiment 3). This finding is replicated in our results. So, although neuroscience is perceived as a more prestigious subject than AI and educational psychology, this perception does not extend to public judgements about the credibility of research evidence framed in neuroscience.

As argued earlier in this paper and the literature, the recipient of the information must be considered as an influencing factor on the way that information is processed and perceived (Metzger and Flanagin 2015). Therefore, the last research question was about the individual differences in the credibility of evidence framed in different disciplines: neuroscience, educational psychology and AI. We investigated all the potential confounding variables we collected to see if any of them would account for the public's bias in their judgments of research credibility. We found significant positive correlations between the familiarity with the article topic and the bias against credibility ($r = 0.374$, $p < 0.0001$). Similar results are frequently cited in the previous literature (Rhodes et al. 2014; Scurich and Shniderman 2014; Im et al. 2017). As argued in the context of scientific reasoning more generally, people judge the credibility of arguments that they are familiar with higher than the credibility of arguments that are unfamiliar to them, or that they disagree with (Greenhoot et al. 2004). For potential reasons behind such bias please see studies on confirmation bias (Nickerson 1998) and availability heuristic (Tversky and Kahneman 1973).

Our results show that after controlling for participants' familiarity with the topic, evidence in the AI framed educational evidence is still considered as statistically significantly less credible than neuroscience and educational psychology frames. The only significant predictor of the incredible AI effect for the overall research credibility was the public attitude towards whether they considered educational psychology researchers adhered to scientific methods or not. Although explaining a relatively small portion of the variance in the composite index, AI frame's relative low credibility to neuroscience and educational psychology is statistically significantly affected by people's attitudes towards educational psychology researchers' adherence to scientific methods. Based on the direction and significance of the coefficient, those people who think that educational psychology researchers adhere to the scientific methods are more likely to judge educational research evidence we presented in our conditions as more credible. In addition to this, two other factors repetitively contributed to sub-categories of our credibility measure were:

(1) the attitude that society views educational psychology as a prestigious discipline (green);

(2) the attitude that AI research can help us understand how children learn (brown).

Interestingly, the demographic features including participants' age, gender, education level, and their academic major were not found to predict the main study effect on the composite credibility score. The only demographic feature which was found to significantly predict the empirical nature and helps with understanding scores was the participants' age. An increase in participants' age leads to a statistically significant decrease in their judgment of the *empirical* and *helps with their understanding* scores of the research credibility.

Limitations and Future Research

First of all, it is important to qualify our sampling of the public studied here. Clearly, due to pragmatic reasons, it is an impossible task to obtain a cross-section of the population that can be referred to as the general public. In Fig. 1 and its following paragraph, we clearly qualify the participants recruited in this research study. Our findings should be interpreted by taking these qualifications into considerations. Moreover, the recruited participants were from an online survey platform who are self-selected people motivated to take part in the survey (Paolacci and Chandler 2014). It is safe to assume that the sample is computer literate, to some extent at least and, some of them at least, might feel more comfortable with what they know about AI. Therefore, there might be a potential bias amongst them against AI, in comparison with the other two subjects. To avoid a potential expertise bias we controlled for the sample's major subject but nevertheless, they might feel proficient in one area, whereas not necessarily in others.

It is important to emphasise that, our demographics show the expected diversity from a public sample (Fig. 1) and sampling from an online survey platform is still considered as more representative compared to sampling from schools, universities, or social media approaches (Casler et al. 2013). However, the results of this paper should be interpreted cautiously with regard to any potential bias of the online recruitment approach we used and further research investigations of the different groupings as well as *how* people consider subject frames when evaluating the credibility of a claim should be undertaken. Moreover, in this study, we collected data from participants familiarity with the subject, attitudes towards the subjects, their demographic features and used this data in our analysis of the results and their interpretation. However, participants' prior knowledge of the topic or their reasoning ability was not explicitly measured in this study and should be considered in future studies. We also did not collect data on participants' reading times which can provide valuable information for the interpretation of the results. Furthermore, the regression predicting the composite credibility variable in Table 4 shows a small portion of its variance explanation, relative to the explained variance of the individual dependent variables. This might suggest that future studies might evaluate more factors to this unidimensional construct. Alternatively, we also suggest rethinking the assumed unidimensionality of the five items. This effect might be 'blurred out' by the weighted averaging of the composition (which might be causing the individual effects to cancel out on each other). Besides, only 3.5% of the multivariate variance was found to be accounted for the group differences in our MANCOVA, which might be lowered by the transformation and averaging of the dependent variables. As this is the first study to show the "in-credible AI effect", we suggest that the results should be approached by caution and argue that further research is required to replicate the findings of this paper and better define different heuristics used by recipients for their credibility scores. For instance, a follow-up study can explore if the findings would be replicated when the research evidence presented to participants are for engineering or general educational interventions rather than for learning as studied in this paper. Moreover, there should be further examinations of whether the lack of perceived credibility judgements of participants would transfer to influencing their choices of intervention in educational settings.

Brief Suggestions of Actions for the AIED Community

There might be various reasons that might lead to the results we presented in this paper including the perceptual processing hypothesis, the potential prestige of the framing discipline, the jargon of certain disciplines might create an impression that the information presented is more or less credible, or the brain-as-engine-of-mind hypothesis. In addition to these, there might also be issues with regard to participants confusion on the relatively unfamiliar nature of AI framing to the contexts of education. It might be the case that AI is only considered as an approach to building tools that can mimic or replicate humans, yet not as a discipline that can be a source of evidence with regard to teaching and learning. Therefore, an AI frame of a psychology finding can just be confusing for the readers and can lead them to score its credibility low. It might also be the case that the participants think that AI in Education is only about *building* systems to optimise learning by changing multiple variables in engineering design solutions. This positioning is different than considering AI as an approach to engage with scientific investigations of learning with hypothesis testing. This could have also potentially explained the findings of this paper.

We also hypothesise that, at least partially, the *in*-credible AI effect might stem from the discrepancy between the actual AI research and its image in popular media. There appears to be a clear lack of public understanding of the mechanisms behind AI in Education. Although a detailed discussion on potential mitigation of the effect observed in this study is not within the scope of this paper, here we suggest that AI in Education community should attempt to be more actively engaged with the public to help mitigate such an effect. Providing training opportunities on the basics of AI might mitigate the adverse effect of AI framing (see, for instance, https://www.elementsofai.com). Moreover, currently, in the field there appears to be a lack of engagement by academics with public, practitioners and developers to provide accessible research evidence and guidance; academics should see this as a worthy enterprise and for which academic staff should be given time and encouragement. There is a lack of systematically reviewed evaluation reports of AI technologies and most independent AI in Education research is not accessible to the key stakeholders of educators and the public. There should be further attempts to create shared understanding opportunities among the key stakeholders of AI in Education community (i.e. www.educate.london).

Conclusions

This study investigated the impact of framing educational research evidence in different subject disciplines upon people's perceptions of the presented research evidence's credibility. We discovered that when educational research evidence is framed within AI, it is considered as less credible compared to when it is framed within neuroscience or educational psychology. The effect is still evident even when the subjects' familiarity with the topic is controlled for. The effect is also evident among educators took part in this study. To the best of our knowledge, this is the first study to show that an AI framing of educational research evidence influences the credibility of presented findings. We aim to study the effect further in both qualitative and quantitative studies, but also with different samples of participants including teachers, researchers, and students who are expected to adopt and use AI technologies in education.

Funding Information This research was partially funded by UCL's Grand Challenge of Transformative Technology small grants award.

Appendix: Data Collection Tools

A. Demographics

Age
Gender
Level of Educational Attainment
Academic major

B. Items measuring attitude towards educational psychology, neuroscience, AI research.

Please rate your level of agreement with the following statements.

1. ***Educational psychology research*** can help us understand how children ***learn*** in school.

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

2. ***Educational psychology researchers*** *adhere to the scientific methods*

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

3. ***Educational psychology,*** *as viewed by society, is a prestigious discipline*

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

4. ***Neuroscience research*** can help us understand how children ***learn*** in school.

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

5. ***Neuroscience researchers*** *adhere to the scientific methods*

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

6. ***Neuroscience,*** *as viewed by society, is a prestigious discipline*

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

7. ***Artificial Intelligence research*** can help us understand how children ***learn*** in school.

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

8. ***Artificial Intelligence researchers*** *adhere to the scientific methods*

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

9. ***Artificial Intelligence,*** *as viewed by society, is a prestigious discipline*

	1	2	3	4	5	6	7	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

C. Complete set of stimuli used in the experiment

Educational Topic	Introduction	Psychological fineding	Frame		
Extraneous verbal input	Extraneous neuroscience	Extraneous AI			
1. Spacing Effect	Researchers investigated the effect of study schedule – the order in which items are studied – on learning. In this study, participants learned paired-associates presented according to one of two schedules. Each paired-associated combined a common word from English with a word from an artificial language (e.g., *dog* – *wug*). Participants in the massed condition learned each paired-associates four times in a row during each trial. Participants in the spaced condition learned each paired-associated four times spread across for different trials.	One day later, participants were given a delayed cued-recall tests. They were presented the first word of each paired-associate and had to recall the second word. The results showed that participants in the spaced condition had better memory for the paired-associates than participants in the massed condition.	A number of theoretical constructs have been proposed to account for the beneficial effects of spacing on memory for repeated items. One popular account sees subjects as devoting less attention or rehearsal to successively repeated, or massed, items. This could be either a consciously controlled strategy of allocating more rehearsal to weaker items or an automatic process that responds to the novelty of a just-presented item.	These results are consistent with the results of a recent neuroimaging study. This study found that participants who spaced their learning showed greater activation in the left frontal operculum than participants who massed their learning. This brain area is thought to be involved in encoding new information via verbal maintenance rehearsal.	These results are consistent with the results of a recent artificial intelligence in education study. Using machine learning techniques and AI approaches, researchers designed an algorithm for providing spaced learning to students. They found that students in the spaced learning condition learned more than the traditional condition.
2.Multitasking	Researchers investigated the effect of media multitasking during learning on academic performance. The participants in this study were students enrolled in an Educational Psychology	At the end of the lecture, both groups took the same quiz on the content of the lecture. The results showed that participants in the experimental group, who multitasked during the	Multitasking refers to the concurrent processing of two or more tasks through a process of context switching. The growth and expansion of communication technology	These results are consistent with the results of a recent neuroimaging study. This study found that participants who reported engaging in more media multitasking during their	These results are consistent with the results of a recent artificial intelligence in education study. Using machine learning techniques and AI approaches, researchers

(continued)

	course. During a lecture, participants in the experimental condition were allowed to multitask – to browse Facebook, text, email, and surf the web. By contrast, participants in the control condition were not allowed to multitask – they were forced to turn off all of their electronic devices.	lecture, scored worse than participants in the control condition, who did not.	have created a multitasking generation of students who believe they are utilizing time more effectively by performing two or more tasks simultaneously.	daily lives had a lower density of neurons in the anterior cingulate cortex than participants who reported engaging in less media multitasking. This brain area is thought to be involved in managing conflicts during task performance.	designed an algorithm for calculating multitasking on their electronic devices during lectures. They found that students who multitask score worse than students who do not multitask.
3. Collaboration	Researchers investigated the better collaboration behaviours of groups of learners. It was hypothesised that observers infer a shared psychological state from synchronized movement rhythms, influencing their attributions of rapport, which in turn affect their collaboration outcomes. In this experimental study movement rhythms of participants were manipulated between participants or kept constant while the source of the emerging movement synchrony is manipulated.	The findings supported the hypothesis that movement synchrony increases attributed rapport. Furthermore, the effects of movement synchrony on perceived unity are not purely perceptual in nature, but caused by psychological inferences. Observers infer the degree to which individuals are a social unit from their movement rhythms. Therefore, more collaborative groups appear to be more synchronized in their behaviours.	The findings supported the hypothesis that movement synchrony increases attributed rapport. Furthermore, the effects of movement synchrony on perceived unity are not purely perceptual in nature, but caused by psychological inferences. Observers infer the degree to which individuals are a social unit from their movement rhythms. Therefore, more collaborative groups appear to be more synchronized in their behaviours.	These results are consistent with the results of a recent neuroimaging study. This study found that high synchrony group members were associated with greater activation in the ventral cingulate cortex, an area associated with emotional processing and ability to collaborate.	These results are consistent with the results of a recent artificial intelligence in education study. Using machine learning techniques and AI approaches, researchers designed an algorithm for calculating synchrony in a remote collaboration setting. By using the distance between the gaze of the emitter and the receiver; they found that if there is more synchrony, collaboration outcomes are better.

(continued)

4. Engagement	Researchers investigated the extent to which student engagement is associated with experimental and traditional measures of academic performance. The sample consisted of 1058 students at 14 four-year colleges and universities that completed several instruments which measured their engagement levels and their academic performances.	Many measures of student engagement were linked positively with such desirable learning outcomes as critical thinking and grades. The results showed statistically significant positive correlations between student engagement results and academic achievement scores, both before and after controls were added for a host of student characteristics. Level of academic challenge, supportive campus climate, reading and writing, quality of relationships, institutional emphases on good practices, and general education gains were some of the control variables tested.	When well-crafted, student surveys can provide insights into the student experience of engagement that other sources of information cannot, such as estimates of one's ability to interact effectively with others on an individual basis or in small groups, and the degree to which one's values and ethics have developed since starting school or college.	These results are consistent with the results of a recent neuroimaging study. This study found that high engagement ratings were associated with greater activation in the substantia nigra / ventral tegmental area. This brain area is thought to be involved in motivation, specifically in learning high-value information.	These results are consistent with the results of a recent artificial intelligence in education study. Using machine learning techniques and AI approaches, researchers designed an algorithm for detecting students' engagement behaviours as part of an intelligent tutoring system. The results show that students' academic achievement is higher when they present higher engagement values.

D. The educational article examples

Example Item 1. Spacing Effect

Introduction

Researchers investigated the effect of study schedule – the order in which items are studied – on learning. In this study, participants learned paired-associates presented according to one of two schedules. Each paired-associated combined a common word from English with a word from an artificial language (e.g., *dog* – *wug*). Participants in the massed condition learned each paired-associates four times in a row during each trial. Participants in the spaced condition learned each paired-associated four times spread across for different trials.

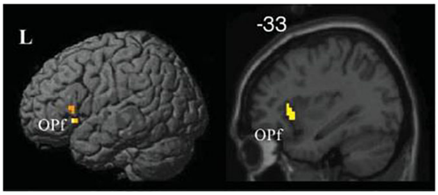
I. Psychological finding alone

One day later, participants were given a delayed cued-recall tests. They were presented the first word of each paired-associate and had to recall the second word. The results showed that participants in the spaced condition had better memory for the paired-associates than participants in the massed condition.

II. Extraneous neuroscience finding

These results are consistent with the results of a recent neuroscience study. This study found that participants who spaced their learning showed greater activation in the *left frontal operculum* than participants who massed their learning. This brain area is thought to be involved in encoding new information via verbal maintenance rehearsal.

IV. Brain Image



Conclusion

These findings suggest that spacing (or "distributing") practice over a long period of time results in better learning than massing (or "cramming") practice over a short period of time.

1a. Spacing Effect

Introduction

Researchers investigated the effect of study schedule – the order in which items are studied – on learning. In this study, participants learned paired-associates presented according to one of two schedules. Each paired-associated combined a common word from English with a word from an artificial language (e.g., *dog* – *wug*). Participants in the massed condition learned each paired-associates four times in a row during each trial. Participants in the spaced condition learned each paired-associated four times spread across for different trials.

I. Psychological finding alone

One day later, participants were given a delayed cued-recall tests. They were presented the first word of each paired-associate and had to recall the second word. The results showed that participants in the spaced condition had better memory for the paired-associates than participants in the massed condition.

II. Extraneous verbal input	A number of theoretical constructs have been proposed to account for the beneficial effects of spacing on memory for repeated items. One popular account sees subjects as devoting less attention or rehearsal to successively repeated, or massed, items This could be either a consciously controlled strategy of allocating more rehearsal to weaker items or an automatic process that responds to the novelty of a just-presented item.
III. Extraneous Image	
Conclusion	These findings suggest that spacing (or "distributing") practice over a long period of time results in better learning than massing (or "cramming") practice over a short period of time.

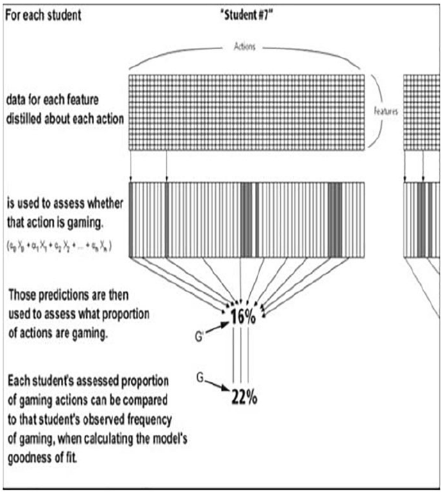
References

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review*, *12*, 159-164. doi:10.3758/BF03196362 [Educational psychology]

Callan, D. E., & Schweighofer, N. (2010). Neural correlates of the spacing effect in explicit verbal semantic encoding support the deficient-processing theory. *Human Brain Mapping*, *31*, 645-659. doi: 10.1002/hbm.20894 [neuroscience]

Example Item 2. Engagement

Introduction	Researchers investigated the extent to which student engagement is associated with experimental and traditional measures of academic performance. The sample consisted of 1058 students at 14 four-year colleges and universities that completed several instruments which measured their engagement levels and their academic performances.
I. Psychological finding alone	Many measures of student engagement were linked positively with such desirable learning outcomes as critical thinking and grades. The results showed statistically significant positive correlations between student engagement results and academic achievement scores, both before and after controls were added for a host of student characteristics. Level of academic challenge, supportive campus climate, reading and writing, quality of relationships, institutional emphases on good practices, and general education gains were some of the control variables tested.
II. Extraneous AI finding	These results are consistent with the results of a recent artificial intelligence in education study. Using machine learning techniques and AI approaches, researchers designed an algorithm for detecting students' engagement behaviours as part of an intelligent tutoring system. The results show that students' academic achievement is higher when they present higher engagement values.

IV. AI's architecture	 For each student "Student #7" Actions Features data for each feature distilled about each action is used to assess whether that action is gaming. $(a_0 X_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n)$ Those predictions are then used to assess what proportion of actions are gaming. 16% G' G 22% Each student's assessed proportion of gaming actions can be compared to that student's observed frequency of gaming, when calculating the model's goodness of fit
Conclusion	These findings suggest that students' academic performance is higher when they present higher engagement results.

2a. Engagement

Introduction	Researchers investigated the extent to which student engagement is associated with experimental and traditional measures of academic performance. The sample consisted of 1058 students at 14 four-year colleges and universities that completed several instruments which measured their engagement levels and their academic performances.
I. Psychological finding alone	Many measures of student engagement were linked positively with such desirable learning outcomes as critical thinking and grades. The results showed statistically significant positive correlations between student engagement results and academic achievement scores, both before and after controls were added for a host of student characteristics. Level of academic challenge, supportive campus climate, reading and writing, quality of relationships, institutional emphases on good practices, and general education gains were some of the control variables tested.
II. Extraneous verbal input	When well-crafted, student surveys can provide insights into the student experience of engagement that other sources of information cannot, such as estimates of one's ability to interact effectively with others on an individual basis or in small groups, and the degree to which one's values and ethics have developed since starting school or college.

IV. Extraneous image	
Conclusion	These findings suggest that students' academic performance is higher when they present higher engagement results.

References

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in higher education*, *47*(1), 1-32. **[Educational psychology].**

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted eInteraction*, *18*(3), 287-314. **[AI in Education]** *Modeling and User-Adapted Interaction*, *18*(3), 287-314. **[AI in Education]**

E. Items measuring the familiarity and credibility of each educational article

Please read the following four brief articles and answer the six questions about each one.

1. Are you familiar with the above topic?
   1) Yes. 2) No.

2. The article was well written.

	1	**2**	**3**	**4**	**5**	**6**	**7**	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

3. The article helped me understand the topic.

	1	**2**	**3**	**4**	**5**	**6**	**7**	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

4. The scientific arguments in the article made sense.

	1	**2**	**3**	**4**	**5**	**6**	**7**	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

5. The article offered strong empirical evidence for its conclusions.

	1	**2**	**3**	**4**	**5**	**6**	**7**	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

6. Do you agree or disagree with the conclusions of the article?

	1	**2**	**3**	**4**	**5**	**6**	**7**	
Strongly disagree	□	□	□	□	□	□	□	**Strongly agree**

References

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly, 93*(1), 59–79. https://doi.org/10.1177/1077699015606057.

Ariely, D., & Berns, G. S. (2010). Neuromarketing: The hope and hype of neuroimaging in business. *Nature Reviews Neuroscience, 11*, 284–292. https://doi.org/10.1038/nrn2795.

Brynjolfsson, E., & Mcafee, A. (2014). The second machine age: Work, Progress, and prosperity in a time of brilliant technologies, MIT Press.

BSA Intelligence (2015). One in three believe that the rise of artificial intelligence is a threat to humanity.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olsen, & C. P. Herman (Eds.), *Social influence: The Ontario symposium* (pp. 3–39). Hillsdale: Erlbaum.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behavior research methods, 46(1), 112-130

Cukurova, M., Luckin, R., & Clark-Wilson, A. (2019). Creating the golden triangle of evidence-informed education technology with EDUCATE. *British Journal of Educational Technology, 50*(2), 1–22.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine, 24*(9), 1342.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). *SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours.* arXiv preprint arXiv: 1704.05972.

Dietterich, T. G., & Horvitz, E. (2015). Rise of concerns about AI: Reflections and directions. *Communications of the ACM, 58*(10), 38–40.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114.

du Boulay, B. (2016). Recent meta-reviews and meta–analyses of AIED systems. *International Journal of Artificial Intelligence in Education, 26*(1), 536–537.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics, 6*, 241–252.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision making, 7*, 746–749.

European Commission (2015) *Special Eurobarometer 427*, Autonomous Systems, June 2015. http://ec.europa.eu/public_opinion/archives/ebs/ebs_427_en.pdf.

Farah, M. J., & Hook, C. J. (2013). The seductive allure of "seductive allure". *Perspectives on Psychological Science, 8*(1), 88–90.

Fast, E., & Horvitz, E. (2016). Identifying dogmatism in social media: Signals and models. *arXiv preprint arXiv:1609.00425*.

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Fernandez-Duque, D., Evans, J., Christian, C., & Hodges, S. D. (2015). Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience, 27*(5), 926–944.

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher, 46*(6), 329–334.

Frey, C. B., & Osborne, M. (2013). The future of employment. *How susceptible are jobs to computerisation.* Published by the Oxford Martin Programme on Technology and Employment.

Gaines-Ross, L. (2016). What do people – Not techies, not companies – Think about artificial intelligence? *In Harvard Business Review,* (24 October 2016).

Giattino, C. M., Kwong, L., Rafetto, C., & Farahany, N. A. (2019). The seductive allure of artificial Intelligence-powered Neurotechnology. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp 397-402, https://doi.org/10.1145/3306618.3314269.

Greenhoot, A. F., Semb, G., Colombo, J., & Schreiber, T. (2004). Prior beliefs and methodological concepts in scientific reasoning. *Applied Cognitive Psychology, 18*, 203–221. https://doi.org/10.1002/acp.959.

Gruber, D., & Dickerson, J. A. (2012). Persuasive images in popular science: Testing judgments of scientific reasoning and credibility. *Public Understanding of Science, 21*(8), 938–948.

Hall, D. W., & Pesenti, J. (2017). Growing the artificial intelligence industry in the UK. *Independent review for the Department for Digital, Culture, Media and Sport/Department for Business, Energy and Industrial Strategy,* https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk.

Hardiman, M., Rinne, L., Gregory, E., & Yarmolinskaya, J. (2012). Neuroethics, neuroeducation, and classroom teaching: Where the brain sciences meet pedagogy. *Neuroethics, 5*, 135–143 https://doi.org/10.1007/s12152-011-9116-6.

Hassabis, D., et al. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron, 95*(2), 245–258.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164. https://doi.org/10.1177/014662168500900204.

Hook, C. J., & Farah, M. J. (2013). Neuroscience for educators: What are they seeking, and what are they finding? *Neuroethics, 6*, 331–341. https://doi.org/10.1007/s12152-012-9159-3.

Im, S. H., Varma, K., & Varma, S. (2017). Extending the seductive allure of neuroscience explanations effect to popular articles about educational topics. *British Journal of Educational Psychology, 87*(4), 518–534.

Keehner, M., Mayberry, L., & Fischer, M. H. (2011). Different clues from different views: The role of image format in public perceptions of neuroimaging results. *Psychonomic Bulletin & Review, 18*(2), 422–428.

Keil, F. C., Lockhart, K. L., & Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *Journal of Experimental Psychology: General, 139*(1), 1.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97.

Malone, T. W. (2018). How human-computer 'Superminds' are redefining the future of work. *MIT Sloan Management Review, 59*(4), 34–41.

Manikonda, L., & Kambhampati, S. (2018). Tweeting AI: Perceptions of lay versus expert Twitterati. In *Twelfth International AAAI Conference on Web and Social Media*.

McCabe, D., & Castel, A. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*, 343–352. https://doi.org/10.1016/j.cognition.2007.07.017.

Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment online. *The handbook of the psychology of communication technology, 32*, 445.

Michael, R. B., Newman, E. J., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non) persuasive power of a brain image. *Psychonomic Bulletin & Review, 20*, 720–725. https://doi.org/10.3758/s13423-013-0391-6.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872. https://doi.org/10.1016/j.jesp.2009.03.009.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*, 184–188. https://doi.org/10.1177/0963721414531598.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology, 19* (pp. 123–205). New York: Academic Press.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. Consciousness and cognition, 8(3), 338–342.

Rhodes, R. E., Rodriguez, F., & Shah, P. (2014). Explaining the alluring influence of neuroscience information on scientific reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1432–1440. https://doi.org/10.1037/a0036844.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*, 437. https://doi.org/10.1038/470437a.

Schwartz, D. L., Blair, K. P., & Tsang, J. M. (2012). How to build an educational neuroscience: Two approaches with concrete instances. *British Journal of Educational Psychology Monograph Series II: Part 8—Educational Neuroscience, 1*, 9–27.

Schweitzer, N. J., & Saks, M. J. (2011). Neuroimage evidence and the insanity defense. *Behavioral Sciences & the Law, 29*, 592–607. https://doi.org/10.1002/bsl.995.

Schweitzer, N. J., Baker, D. A., & Risko, E. F. (2013). Fooled by the brain: Re-examining the influence of neuroimages. *Cognition, 129*, 501–511. https://doi.org/10.1016/j.cognition.2013.08.009.

Scurich, N., & Shniderman, A. (2014). The selective allure of neuroscientific explanations. *PLoS One, 9*, 1–6. https://doi.org/10.1371/journal.pone.0107529.

Shariff, S. M., Zhang, X., & Sanderson, M. (2017). On the credibility perception of news on twitter: Readers, topics and features. *Computers in Human Behavior, 75*, 785–796.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... & Leyton-Brown, K. (2016). Artificial Intelligence and Life in 2030. One hundred year study on artificial intelligence: Report of the 2015–2016 study panel. Stanford University, Stanford, http://ai100.stanford.edu/2016-report.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Boston: Allyn and Bacon.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232. https://doi.org/10.1016/0010-0285(73)90033-9.

Uebersax, J.S. (2006). Likert scales: Dispelling the confusion. Statistical Methods for Rater Agreement website.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Wassmer, M., & Eastman, C. M. (2005). Automatic evaluation of credibility on the web. *Proceedings of the American Society for Information Science and Technology, 42*(1).

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*, 470–477. https://doi.org/10.1162/jocn.2008.20040.

Weisberg, D. S., Taylor, J. C., & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision making, 10*, 429–441.

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Socety A, 376*(2133), 20180085.

Yang, C. Y. D., Ozbay, K., & Xuegang, J. B. (2017). Developments in connected and automated vehicles. *Journal of Intelligent Transportation Systems, 21*(4), 251–254. https://doi.org/10.1080/15472450.2017.1337974 Minutes. 2016. *60 minutes poll*: Artificial intelligence.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.