Vignette Marking Scheme (Studies 2 and 3)

Condition	Abbreviation	Presenting Complaint
Temporal Arteritis	TA	Patient is a 68 year old male presented with fever and arthralgi
Ulcerative Colitis	UC	Patient is a 60 year old male presented with 2 day history of bl
Miliary Tuberculosis	MTB	Patient is a 62 year old male admitted for fevers and generalise
Aortic Dissection	AD	Patient is a 58 year old female presented with shortness of brea
Guillain-Barré Syndrome	GBS	Patient is a 67 year old male presented with weakness of the leg
Thrombotic Thrombocytopenic Purpura	TTP	Patient is a 20 year old male was admitted from an outside hos

Table S1: Marking scheme used to denote differentials that are considered as correct for each of the six patient cases/vignettes. The same marking scheme is applied for online and think-aloud vignette studies. The presenting complaint is shown to participants at the start of the case, before they start seeking information.

Vignette Information Requests

Patient History	Physical Examinations	Testing
History of Presenting Complaint	Take Pulse	Urine Dipstick
Past Medical History	Measure Blood Pressure	ECG
Medications	Assess Respiratory Rate	Abdominal CT Scan
Allergies	Auscultate Lungs	Venous Blood Gas
Family History	Auscultate the Heart	CRP and ESR
Social History	Assess Eyes	Clotting Test
	Measure Temperature	FBC
	Abdomen Examination	Other Biochemistry tests
	Rectal Examination	UREA and Electrolytes
	Neck/Throat Examination	Chest X-Ray
	Assess Head	
	Neurological Exam Record	
	Assess Extremities	

Table S2: Full list of possible information requests that participants can make. This set of information is the same for all cases. The same vignettes and corresponding information are used for the online and think-aloud vignette studies.

Calibration of Confidence to Alternative Accuracy Measures

Differential Accuracy

When comparing Differential Accuracy (if a correct differential is provided in the participant's list) to Confidence, we find, across stages, participants' Confidence was not aligned to their Accuracy. Instead, we find evidence of underconfidence at all stages. There was evidence of a significant difference between the two at the Patient History ($t(84) = 8.24$, MDiff = 0.24, $p < .001$), Physical Examination stage ($t(84) = -9.09$, MDiff = -0.25, $p < .001$), and Testing stage ($t(84) = -7.74$, MDiff = -0.22, $p < .001$).

In order to examine the observed underconfidence in more detail, we compare confidence and Differential Accuracy by case (the mean values of which can be found in Table 1 of the main thesis). We conducted paired t-tests for each condition's cases by comparing Differential Accuracy and confidence values (at the final Testing stage) to observe if they significantly differ from each other. A p value of less than .05 is interpreted as evidence for overconfidence or underconfidence (depending on the direction of the effect). We observed underconfidence for the GBS case ($t(84) = -7.43$, MDiff = -0.39, $p = < .001$), the TA case ($t(84) = -5.07$, MDiff = -0.25, $p = < .001$), the TTP case ($t(84) = -3.23$, MDiff = -0.2, $p = < .001$) and the UC case ($t(82) = -14.83$, MDiff = -0.38, $p = < .001$). The remaining cases did not yield a significant effect.

Highest Likelihood Accuracy

When comparing Highest Likelihood Accuracy (likelihood assigned to the highest likelihood differential if it is correct) to Confidence, we find, across stages, participants' Confidence was not aligned to their Accuracy. Instead, we find evidence of overconfidence at all stages. There was evidence of a significant difference between the two at the Patient History ($t(84) = -2.49$, MDiff = -0.05, $p = 0.01$), Physical Examination stages ($t(84) = 4.45$, MDiff = 0.09, $p < .001$), and Testing stage ($t(84) = 6.84$, MDiff = 0.16, $p < .001$).

In order to examine the observed overconfidence in more detail, we compare confidence and Highest Likelihood Accuracy by case (the mean values of which can be found in Table 1 of the main thesis). We conducted paired t-tests for each condition's cases by comparing Highest Likelihood Accuracy and confidence values (at the final Testing stage) to observe if they significantly differ from each other. A p value of less than .05 is interpreted as evidence for overconfidence or underconfidence (depending on the direction of the effect). We observed overconfidence for the AD case ($t(84) = 8.92$, MDiff = 0.37, $p = < .001$), the MTB case ($t(83) = 7.66$, MDiff = 0.35, $p = < .001$) and the TTP case ($t(84) = 4.09$, MDiff = 0.21, $p = < .001$). The remaining cases did not yield a significant effect.

Debrief Questionnaire from Think-Aloud Study

Each question has a corresponding follow-up question below in case they are not answered by responses to the main questions.

- 1. What's your general approach to making diagnoses? *Follow-Up:* Do you have those cognitive aids or frameworks you use?
- 2. Do you tend to keep a broad set of differentials in mind? *Follow-Up:* Are there particular situations where having a narrower set would be more useful?
- 3. How do you decide what information or tests to get on a patient? *Follow-Up:* Would you say you tend to seek information to confirm or to rule out differentials that you have in mind?
- 4. How similar was your diagnostic reasoning on this task versus how you would approach diagnosis in real life? *Follow-Up:* Was there anything that prevented you from approaching the task as you would in real life?

Diagnostic Appropriateness Marking Scheme for VR Study

The table below shows differentials for each scenario that were categorised as probable/possible and those categorised as improbable/unlikely. Any differentials provided by participants that were not included in this table were considered incorrect.

Scenario	Probable/Possible Differentials	Improbable/Unlikely Differentials
Asthma	Asthma / asthma excerbation Pneumonia / LRTI RSV / Viral URTI Foreign Body Anaphylaxis Viral Induced Wheeze	Epiglotitis Croup PE
DKA	DKA URTI / throat infection / tonsillitis Gastroenteritis / abdominal infection Insulin non compliance Sepsis Viral infection	Alcohol ingestion Sickle Cell Inborn errors of metabolism
Seizure	Epilepsy / Febrile Seizure Meningitis / CNS infection / encephalitis Hypo / hypoglycaemia Non accidential injury (NEA) Space occupying lesion (SOL) / tumour	Fictitious / malingering Alcohol withdrawing Sickle cell Inborn errors of metabolism
Pneumonia	Pneumonia / LRTI URTI / cold / flu Viral LRTI Asthma Inhaled foreign body (FB)	Anaphylaxis Pleural effusion Pneumothorax

R Environment and Packages

```
# print("R version:")
# version$version.string
#
# print("Rstudio version:")
# rstudioversion <- rstudioapi::versionInfo()
# rstudioversion$version
#
# print("Citations for packages used:")
# get_pkgs_info(pkgs = required_packages, out.dir = getwd())
# pkgs <- scan_packages()
# get_citations(pkgs$pkg, out.dir = getwd(), include.RStudio = TRUE)
# cite_packages(pkgs = required_packages, output = "table", out.format = "Rmd", out.dir = getwd())
```

```
#
# required_packages %>%
#   map(citation) %>%
#   print(style = "text")
```