**Evaluation Methodology for the Protocol**

After developing a threat modeling protocol tailored to horizontal cooperatives, it’s crucial to **evaluate its effectiveness**. The evaluation outlined here is based entirely on the user-provided documents, ensuring we use the metrics and methods the user has in mind. The goal of evaluation is to see how well the new protocol works in practice (especially compared to a standard method like STRIDE) and whether it truly meets the needs of decentralized, non-hierarchical organizations. We will use a combination of **quantitative metrics** (to measure things like how many threats are caught) and **qualitative assessments** (to gauge usability and cultural fit). Below is the consolidated evaluation plan.

**Key Evaluation Metrics and Data Collection**

The performance of the threat modeling protocol will be judged on several key metrics. These metrics cover both the **security outcomes** and the **process fit** for a cooperative context:

* **Usability:** *How easy and engaging is the protocol for participants to use?* This looks at the clarity of steps and the effort required to carry them out. If the process is too cumbersome or confusing, members might disengage.
  + *Data Collection:* After using the protocol in a case study, gather feedback through surveys and interviews. Ask participants questions like “Did you understand what to do at each step?”, “Which parts were frustrating or difficult?”, “Would you be willing to do this process regularly?”. We would note if members felt overwhelmed by jargon or if certain activities (like creating attack trees) needed more guidance.
  + *Success Indicator:* Ideally, participants report that the process was clear enough and that they felt able to contribute. High usability would be reflected by minimal confusion during sessions and positive comments such as “It was a bit long but I learned a lot” rather than “I had no idea what was happening.”
* **Precision of Threat Identification:** *Are the threats identified by the protocol actually relevant and real (true positives) as opposed to far-fetched or incorrect (false positives)?* Precision is defined as the proportion of identified threats that were judged to be valid. For example, if the group identified 20 threats but 5 of those turned out not applicable or duplicates, that’s 15/20 = 75% precision.
  + *Data Collection:* To measure this, we need a reference list of “actual threats” present in the scenario. This could be constructed by an expert panel or by combining the results of multiple methods (say, everything found by either our protocol or STRIDE, assuming between them they found most issues). Each threat found by the new protocol is then checked: is it a real threat to the system? (This can be decided by consensus of facilitators or subject-matter experts in the case study.) We then compute precision = (number of correct threats found by our protocol) / (total threats it found).
  + *Success Indicator:* A higher precision means the protocol didn’t waste people’s time on irrelevant issues. We expect good precision if the structured brainstorming (with Security Cards, etc.) stays focused. However, very high precision at the cost of missing things might be bad – which is why we pair it with recall.
* **Recall of Threat Identification:** *How many of the total actual threats did the protocol manage to find?* In other words, did we miss anything big? Recall is defined as the proportion of actual threats that were identified by the protocol. For example, if there were 20 real threat scenarios in the case, and the protocol found 15 of them, recall = 75%.
  + *Data Collection:* Using the same reference set of actual threats (from expert analysis or union of all methods), count how many of those the new protocol caught. Compute recall = (correct threats found by protocol) / (total real threats). This requires that we have a fairly comprehensive view of the scenario’s threats, which is why we do comparative sessions and expert reviews.
  + *Success Indicator:* A high recall (closer to 100%) is important – missing a critical threat means a blind spot in security. We anticipate that because the protocol encourages broad thinking (including social threats), it might catch things a traditional method might miss (like governance attacks). We’ll specifically check recall on those cooperative-specific threats (Sybil, insider issues) to see if the protocol covers them well.
* **Operational Efficiency:** *How much time and effort does it take to use the protocol?* In a cooperative, people’s time is precious and lengthy processes can be a barrier. We examine how efficiently the protocol can be executed.
  + *Data Collection:* Measure the total time spent on each phase during the case study (e.g., time to map system, time to identify threats, etc.) and how many person-hours were involved (e.g., 10 people in a 2-hour session = 20 person-hours). We’ll also observe if any steps caused long delays or needed multiple meetings. Additionally, note any bottlenecks or points where discussion got stuck. We might also compare the time taken by our protocol vs. time taken by STRIDE in the parallel group, to see if one is significantly more efficient.
  + *Success Indicator:* Efficiency is not just raw speed; it’s also “smoothness.” If the protocol takes a bit longer because more people are involved, that might be acceptable as long as it doesn’t feel like a slog. But if sessions drag on far beyond planned or decisions can’t be reached without marathon meetings, that’s a problem. An efficient process would be one that fits into the organization’s workflow (say, can mostly be done in normal meeting times or short workshops) and doesn’t require heroic effort to complete.
* **Adaptability:** *Can the protocol adapt to changes and different contexts easily?* This metric checks if the process is flexible when the situation changes (which is common in coops). For instance, if mid-evaluation we introduce a new tool or the organization’s structure shifts, does the threat model update gracefully or does it break the process?
  + *Data Collection:* During longer-term evaluation, we might deliberately introduce a change – e.g., simulate the coop adopting a new technology after the initial threat model is done – and see if participants can update the threat model without confusion (“Okay, let’s add that new asset and consider new threats”). Also, over the course of a long-term study (if one coop uses the protocol for months), we track how they incorporate real changes (new members, new services, etc.). Do they run a mini-session to update things, or do they feel they have to start over?
  + *Success Indicator:* The protocol is adaptable if participants manage to revise the model incrementally. For example, if adding a new asset only requires revisiting one or two steps quickly, that’s good. If they find they have to re-do everything from scratch or the framework doesn’t accommodate a new type of threat well, that’s a sign of low adaptability.
* **Scalability:** *Does the protocol work for different sizes of groups and complexity without breaking down?* We assess if a larger cooperative (or a more complex system) can use the protocol as effectively as a small one. This is important because some methods work fine for a small project but don’t scale to big teams or complicated, multi-part systems.
  + *Data Collection:* Test the protocol in at least two environments: e.g., a small coop of 5-10 people and a larger one of 50+, or on a simple system vs. a complex system with many components. Compare outcomes like the number of threats identified, the participation rate (% of members contributing), and the time taken. Also, gather feedback from participants in the larger scenario: did the process still feel manageable? Or did it become unwieldy with so many voices?
  + *Success Indicator:* A scalable protocol might need minor modifications (like using breakout groups for a big coop, which our protocol already anticipates). If, with those adaptations, the larger group still successfully completes the process and feels it was useful, that indicates scalability. We also look at whether the threat list or diagrams became too complex to handle – if participants in a large coop say “this was just too chaotic,” that’s a scalability issue. Efficiency ties in here: maybe the process remained effective but took much longer with more people – we’d note that as a scaling cost.
* **Preservation of Horizontality (Democratic Centralism):** *Does the protocol preserve the cooperative’s horizontal character and participatory decision-making?* This is a qualitative but vital metric: we want to ensure that by introducing a structured process, we haven’t inadvertently created hierarchy or excluded voices. We also want to see if the protocol actually **strengthens** democratic participation in security (as intended) and if any temporary centralization (for efficiency or emergency) stays within acceptable bounds.
  + *Data Collection:* Observe and interview. During sessions, note how decisions are made: Is one facilitator dictating, or are they truly incorporating input? Do certain people dominate (like tech experts) or are others actively participating? After sessions, ask members: “Did you feel the process was inclusive? Did you feel any part of it took power away from the group or from you personally?” Also, if an emergency simulation was done, did the designated incident team act and then report back properly, and did members feel that was done in the spirit of the agreement? We also count participation: what fraction of members took part in the threat modeling and in key decisions? High participation suggests the process was accessible and engaging.
  + *Success Indicator:* The best outcome is participants saying things like “I felt heard,” “It was collaborative and reflected how we normally decide things,” and perhaps even “This process got more of us involved in security than usual.” If the protocol is truly good for horizontality, it might even improve transparency (members learn more about systems) and empowerment (non-tech members feel they can contribute to security). On the other hand, red flags would be comments like “I guess the tech team did most of it and just told us the results,” or “It felt like we set up a committee and then everything happened there, we didn’t really know.” We want to see broad, equitable participation and respect for the agreed rules (e.g., if a decision was made by vote, that it was carried out without someone overriding it). Essentially, measure whether the cooperative values were preserved or enhanced.

Each of these metrics will provide insight into different aspects: **effectiveness** (precision/recall), **efficiency** (usability, time), and **fit** (horizontality, adaptability). The evaluation will gather both numbers (like counts of threats, time taken) and narratives (what users felt and experienced) to give a well-rounded picture.

**Evaluation Design: Comparative Case Studies**

The evaluation will be carried out through a series of case studies in real or simulated cooperative environments. We will compare the new protocol against a baseline (the STRIDE method) to see differences. The general structure is:

1. **Preparation and Training:** We will train participants in the organization on both the new protocol and the baseline (STRIDE) so that lack of understanding isn’t a factor. Everyone will know how to use both methods at a basic level before starting.
2. **Parallel Threat Modeling Sessions:** For each case study (each participating organization):
   * Divide participants into two groups with similar composition (mix of roles/skills in each).
   * Have one group use the **new protocol** and the other use **STRIDE** to analyze the **same system or scenario** independently. For fairness, they should operate under the same conditions/time allotment.
   * For example, if evaluating a tech coop’s communication platform security: Group A applies our cooperative protocol, Group B applies a standard STRIDE process, each for, say, a 2-hour session.
   * Observers (researchers) will monitor both sessions, taking notes on how the discussion flows, how many threats are found, how decisions are made, etc. They won’t interfere, but they will time each section and note any issues (e.g., “Group B (STRIDE) got stuck on understanding ‘repudiation’, Group A (new) spent more time in brainstorming,” etc.).
   * Afterward, collect the outputs: threat lists, any diagrams, and proposed mitigations from both groups.
   * Have a debrief survey or meeting with participants to capture immediate reactions: “Which method did you prefer and why? Did you feel one missed things the other caught? Was one easier?”
3. **Simulated Attack Exercises:** In addition to the modeling sessions, we will conduct a **tabletop simulation** (or actual controlled drill) for certain threats to see how the protocol’s recommendations hold up:
   * This could involve actually testing an aspect, like attempting a mock phishing on both groups (with their knowledge as a test) to see if the threats identified lead to better preparedness.
   * Or simulate an insider trying to do something sneaky during the process to see if group norms catch it (like someone introduces a fake threat to see if the group validates sources).
   * These simulations help evaluate if the protocol has practical effect – e.g., did the coop using the protocol respond faster or more cohesively to a surprise scenario than one that didn’t?
4. **Long-Term Deployment (if possible):** Have at least one cooperative use the new threat modeling protocol as part of their regular operations for a longer period (several months). During this time, track:
   * Do they actually refer to it in meetings?
   * Do they update it when changes happen?
   * Are there any security incidents, and how do they handle them?
   * This will show real-world sustainability and integration.

Throughout these phases, maintain ethical considerations: participation is voluntary, coops consent to the study, data is anonymized if shared, and the process of evaluation itself is transparent to them (in line with their horizontal nature – we treat them as collaborators in research, not subjects).

**Data Analysis and Interpretation**

Once data is collected, we will analyze the results along the metrics:

* Compare the **number and type of threats** identified by the new protocol vs. STRIDE. For example, does our protocol find more social/governance threats than STRIDE does? Does STRIDE find technical threats our protocol missed? This will directly affect precision/recall analysis.
* Analyze the **mitigations proposed** by each – are the protocol’s suggestions more aligned with cooperative practices while STRIDE’s are more technical?
* Look at the **metrics quantitatively:** e.g., Precision of our protocol = X%, Precision of STRIDE = Y%. Same for recall, time taken, etc. This comparative lens will highlight strengths/weaknesses (maybe our protocol takes a bit longer, but finds more threats and is preferred by users).
* Evaluate qualitative feedback: If many participants say the new method was more inclusive and engaging, that’s a success even if, say, it took an extra 30 minutes. We’ll compile common themes from interviews (like “We loved the persona exercise” or “We were confused by the card prompts”).
* Address **challenges observed:** For instance, if in a large coop the process nearly derailed (a challenge of scalability), note that and any solutions they came up with on the fly.
* Check the **horizontality preservation:** Did any group feel sidelined or did power dynamics shift? Ideally, we may see evidence like in one case an outspoken leader had to listen because the structure forced turn-taking – which is a good outcome. Or we may see that without a facilitator the discussion wandered – indicating we need clearer facilitation guidelines but not necessarily a hierarchy.

**Anticipated Challenges in Evaluation**

The documents highlight some specific difficulties we expect and plan to handle:

* **Cultural Fit:** Cooperatives might be skeptical of being “evaluated” at all, or wary of any process that feels corporate. We have to ensure the evaluation respects their culture. For example, we will involve them in designing the scenarios and make sure the process doesn’t violate their norms (no surprise drills without consent, as that could be seen as deceptive).
* **Comparing Apples to Oranges:** Our protocol covers things (like social threats) that STRIDE doesn’t explicitly. When comparing recall, we must carefully define the universe of threats. We might find STRIDE group doesn’t mention a Sybil attack at all – is that a miss for them or just out of scope? We’d likely consider it a miss, which shows our protocol adds value. But we must ensure we’re fair in comparisons.
* **Measuring “No Incident”:** If a coop using the protocol has no security incidents for a year, is it because the protocol worked or just luck? To address this, we rely on simulated incidents and the thoroughness of the threat model rather than waiting for a breach (which is not desirable). We also use participant confidence as a proxy: do they feel safer or more in control? That sentiment, while subjective, indicates the protocol’s impact on perceived security.
* **Volunteer Fatigue:** We must be careful not to overburden coop members with too many exercises, which could lead to drop-off. We will streamline sessions and, if possible, integrate with meetings they already have. And we’ll clearly value their time by sharing back results and making it beneficial (like a free security consultation in practice).
* **Data Sensitivity:** Coops might share internal issues during threat modeling. We ensure confidentiality in research reporting. When we publish or share findings, we might anonymize the cases (“Coop A identified X threats, Coop B identified Y”) unless they’re comfortable being credited.

**Evaluation Outcomes and Use**

By the end of the evaluation, we expect to have:

* A **comparative analysis** showing how the cooperative threat modeling protocol stacks up against a traditional method in terms of thoroughness (precision/recall), engagement (usability, participation), and relevance (the kinds of mitigations and strategies generated).
* **Concrete metrics**: e.g., “Protocol identified 30% more threats overall than STRIDE in similar settings, with particular strength in identifying insider and governance-related threats that STRIDE largely missed. However, it required ~20% more time on average. Participants rated its usability 4.5/5, versus 3.5/5 for STRIDE, citing better collaboration.”
* **Insights on improvements:** The evaluation isn’t only a yes/no verdict; it will feed back into refining the protocol. For instance, if we find participants were confused in Step 4 (Persona creation), we can tweak the instructions for that step in the next iteration of the protocol. Or if certain parts took too long, we can find ways to streamline them without losing value.
* **Evidence of horizontality preservation:** If successful, we’ll have qualitative examples like, “In Coop A’s use of the protocol, even non-technical members contributed multiple threats and felt ownership of the outcomes, whereas in the STRIDE session they mostly stayed silent.” Such evidence would demonstrate the protocol’s advantage in a horizontal context.
* **Documentation of challenges:** e.g., “Large-group facilitation needs more support; we observed in Coop B (50 members) that without breaking into subgroups, the brainstorming became dominated by a few voices. Future protocol versions should explicitly recommend subgroup work for groups larger than 15.” This kind of note ensures the protocol can be adjusted to truly be modular for different scales.

In summary, the evaluation will rigorously test whether this threat modeling protocol truly helps cooperatives improve their security in a democratic, effective way. We will learn not just if it “works,” but how it works in practice, what benefits it brings (like better inclusion or novel threat insights), and what pitfalls to avoid (like potential slowdowns or misunderstandings). By using a mix of metrics and closely observing real-world use, the evaluation will provide a solid foundation for trusting and further developing the protocol, and it will demonstrate to other cooperatives the tangible value of adopting a security process tailored to their needs.